FORMATTING INSTRUCTIONS FOR ICLR 2027 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured generation models are commonly optimized for average performance, although low-frequency semantic events may carry disproportionate importance in high-stakes domains. We study this problem in structured radiology report generation, where strong aggregate text and semantic metrics can coexist with degraded reliability on rare diagnostic findings. We formulate this failure as a semantic-tail robustness problem and introduce Diagnosis-First Robust Training (DFR), which explicitly increases optimization pressure on diagnosis-priority examples and underperforming groups. Our framework uses reference-side semantic representations to characterize abnormality and diagnostic rarity while separating average generation quality from worst-group and rare-finding reliability. Across single-source and cross-source radiology report generation settings, we find that models with similar aggregate quality can exhibit substantially different semantic-tail behavior, and that diagnosis-priority training shifts the operating point between average fidelity and diagnosis-priority reliability. On the primary benchmark, full DFR yields the strongest worst-group operating point, while alternative diagnosis-priority objectives favor Group 11 performance or rare recall. Under the pooled cross-source setting, DFR improves worst-group F1 from 0.836 to 0.867, rare recall from 0.733 to 0.757, and reduces the rare false-negative rate from 0.267 to 0.243 while maintaining competitive aggregate quality. These results show that diagnosis-priority robustness is a distinct property of structured clinical generation that should be evaluated explicitly rather than inferred from average generation metrics alone.

1 INTRODUCTION

Structured generation systems are usually trained and evaluated through objectives that aggregate performance over the full data distribution. This design is natural when average predictive quality is the primary target, but it can be inadequate when low-frequency semantic events carry disproportionate importance. A model can therefore achieve strong average text or semantic similarity while remaining unreliable on a small subset of examples. In high-stakes settings, this distinction matters because the cases that contribute least to an average objective may still contain clinically important abnormalities.

Structured Radiology Report Generation (SRRG) provides a concrete setting in which this mismatch is visible. Radiology report generation has been studied with memory-driven transformers, multimodal alignment, and more recent large-model-based approaches Chen et al. (2022); Yang et al. (2022); Zhou et al. (2024). SRRG further reformulates this task toward standardized structured Impression generation Delbrouck et al. (2025). Existing systems are commonly evaluated with aggregate generation metrics such as BLEU Papineni et al. (2002), ROUGE-L Lin (2004), and BERTScore Zhang et al. (2020), together with clinically motivated semantic metrics extracted from generated and reference reports. Clinical evaluation has increasingly incorporated structured entities, relations, and clinically grounded report-level metrics Jain et al. (2021); Zhang et al. (2024). These metrics are useful for characterizing overall report quality, but they do not directly reveal whether performance is preserved on rare and abnormal findings. A model with competitive aggregate scores may still omit or distort findings concentrated in a small diagnosis-priority subset.

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.

We study this problem as *semantic-tail robustness*. The central question is whether strong aggregate performance in SRRG implies reliable behavior on diagnosis-priority minority slices, and whether training objectives can shift the trade-off between average report quality and reliability on those slices. We define diagnosis-priority groups from reference-side semantic labels using two attributes: abnormality and diagnostic rarity. This yields normal-common, normal-rare, abnormal-common, and abnormal-rare slices. The abnormal-and-rare slice serves as the primary diagnosis-priority group, while worst-group performance over the non-degenerate operative groups provides a broader robustness measure.

This formulation differs from treating rarity as a purely statistical property of examples. In SRRG, the relevant failure is not simply that a case is infrequent, but that a rare semantic finding may be clinically consequential and can be obscured by majority patterns during training. This issue is related to prior work on worst-group and subgroup robustness, which has shown that models optimized for average risk may underperform on minority or difficult groups Sagawa et al. (2020); Liu et al. (2021). We therefore distinguish aggregate generation quality from diagnosis-priority reliability and evaluate both explicitly.

To study whether training can alter this reliability profile, we introduce Diagnosis-First Robust Training (DFR). DFR combines diagnosis-priority weighting with group-aware reweighting. The first component increases the contribution of diagnosis-priority examples, while the second increases the influence of underperforming groups. Rather than assuming that one configuration should improve every metric simultaneously, we analyze DFR as changing the operating point among aggregate quality, worst-group reliability, abnormal-and-rare performance, and rare-finding recall.

We evaluate this framework in three settings: a primary SRRG benchmark based on CheXpert Plus Chambon et al. (2024), a second-source SRRG-lite setting derived from MIMIC-CXR Johnson et al. (2019), and a pooled cross-source regime with a fixed training budget. Across these settings, our results show that aggregate quality and diagnosis-priority robustness can diverge. On the primary benchmark, full DFR improves worst-group F1 from 0.781 to 0.820, while group-only training attains the strongest Group 11 F1 and rare recall. Under the pooled cross-source setting, DFR improves worst-group F1, rare recall, and rare false-negative rate together while maintaining competitive ROUGE-L and semantic F1. These results support an operating-point view of robust structured generation rather than a single aggregate optimum.

Our contributions are threefold:

1. We formulate diagnosis-priority reliability in SRRG as a semantic-tail robustness problem and show that strong aggregate text and semantic performance can coexist with substantially different behavior on abnormal-and-rare and worst-performing slices.
2. We introduce Diagnosis-First Robust Training, which combines diagnosis-priority weighting with group-aware reweighting, and characterize how its components shift the operating point between aggregate report quality, worst-group robustness, and rare-finding reliability.
3. We evaluate diagnosis-priority robustness across primary, second-source, and pooled cross-source settings, showing that robustness profiles depend on data composition and that the pooled regime yields the clearest joint improvement in worst-group F1, rare recall, and rare false-negative rate.

Figure 1 summarizes the overall diagnosis-first robustness framework.

2 Related Work

2.1 Radiology report generation and structured reporting

Radiology report generation has been studied with memory-augmented transformers Chen et al. (2022), multimodal alignment and external knowledge Yang et al. (2022), and more recent large-model-based approaches Zhou et al. (2024); Miao et al. (2025). These approaches have substantially improved the fluency and average-case quality of generated reports. Structured Radiology Report Generation further reformulates the task toward standardized outputs that are easier to compare, evaluate, and integrate into downstream workflows Delbrouck et al. (2025). Our work addresses a

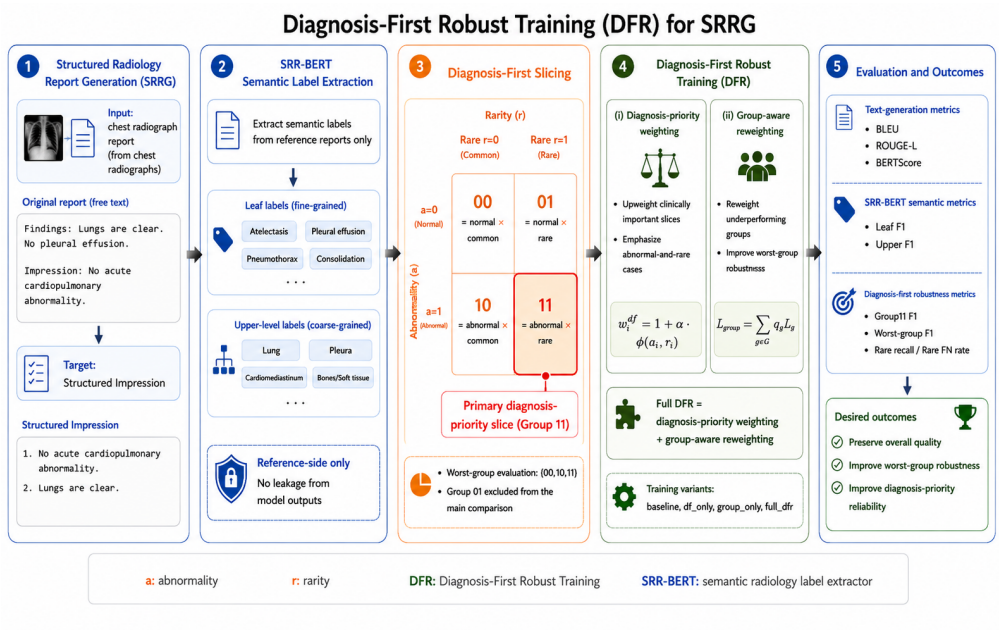


Figure 1: Overview of the diagnosis-first robustness framework for Structured Radiology Report Generation. Reference-side semantic labels define abnormality–rarity slices. Diagnosis-First Robust Training combines diagnosis-priority weighting with group-aware reweighting, while evaluation separates aggregate report quality from diagnosis-priority reliability.

complementary question: whether strong average report quality also implies reliable behavior on clinically important minority slices.

2.2 Clinical evaluation of generated reports

Text-overlap metrics such as BLEU Papineni et al. (2002) and ROUGE-L Lin (2004) remain common in report-generation evaluation, while BERTScore Zhang et al. (2020) provides an embedding-based measure of semantic similarity. Clinical report generation has also motivated entity-level and relation-level evaluation, including RadGraph Jain et al. (2021), as well as public benchmarking efforts such as ReXrank Zhang et al. (2024). These approaches make it possible to evaluate whether clinically meaningful content is preserved even when surface forms differ. We build on this direction by using reference-side semantic labels not only for evaluation but also to define diagnosis-priority slices for robustness analysis.

2.3 Long-tail learning and subgroup robustness

Average-risk training can underperform on minority or difficult groups, motivating group-robust and worst-case optimization approaches Sagawa et al. (2020); Liu et al. (2021). Related long-tail methods address imbalance through reweighting, resampling, or robust objectives. In structured clinical generation, however, the relevant groups should reflect semantic and diagnostic relevance rather than arbitrary partitions alone. Our approach defines groups through abnormality and rarity and evaluates the resulting trade-off between aggregate generation quality and diagnosis-priority reliability.

2.4 Cross-source robustness in medical AI

Medical datasets can differ in reporting style, annotation conventions, patient populations, acquisition protocols, and label prevalence. MIMIC-CXR Johnson et al. (2019) and CheXpert Plus Chambon et al. (2024) provide complementary chest radiography resources with associated reports and metadata. Evaluation on a single source may therefore provide an incomplete picture of reliability.

We use a second-source setting and a pooled cross-source regime to examine whether diagnosis-priority behavior remains visible beyond one benchmark-specific distribution.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of free-text chest radiology reports x_i and structured Impression targets y_i. A sequence model f_θ generates

$$\hat{y}_i = f_\theta(x_i).$$

Let $\ell_i = \ell(f_\theta(x_i), y_i)$ denote the per-example sequence loss. Standard training minimizes the average empirical risk

$$\mathcal{L}_{\text{base}} = \frac{1}{N} \sum_{i=1}^{N} \ell_i.$$

This objective gives each example equal contribution before averaging over the empirical distribution. Consequently, low-prevalence semantic patterns can exert limited influence on the total objective even when their diagnostic importance is high.

3.2 DIAGNOSIS-PRIORITY SLICING

We construct diagnosis-priority slices using semantic labels extracted from the reference Impression. The label extractor is applied only to the reference side when groups are defined, preventing group assignments from depending on model predictions.

For each example i, we define an abnormality indicator $a_i \in \{0, 1\}$ and a rarity indicator $r_i \in \{0, 1\}$. Their combination defines

$$g_i = (a_i, r_i) \in \{00, 01, 10, 11\}.$$

The groups correspond to normal-common (00), normal-rare (01), abnormal-common (10), and abnormal-rare (11). Group 11 is the primary diagnosis-priority slice.

Group 01 is nearly degenerate in all evaluated settings. We therefore exclude it from the main worst-group computation and define worst-group F1 over the operative set

$$\mathcal{G} = \{00, 10, 11\}.$$

Rarity is defined from training data only. We evaluate a predefined threshold grid and use a support-validity rule to avoid thresholds that produce a degenerate abnormal-and-rare slice. We fix $\tau = 2000$ in all main experiments. Validation and test performance are not used to define the rarity threshold.

3.3 DIAGNOSIS-PRIORITY WEIGHTING

Diagnosis-priority weighting increases the contribution of examples in clinically relevant minority slices. Let

$$w_i^{\text{df}} = 1 + \alpha\phi(a_i, r_i),$$

where $\phi(a_i, r_i)$ emphasizes abnormal-and-rare cases and α controls the weighting strength. The diagnosis-priority objective is

$$\mathcal{L}_{\text{df}} = \frac{1}{N} \sum_{i=1}^{N} w_i^{\text{df}} \ell_i.$$

3.4 GROUP-AWARE REWEIGHTING

For each operative group $g \in \mathcal{G}$, let $\mathcal{D}_g$ be the corresponding training subset and define

$$\mathcal{L}_g = \frac{1}{|\mathcal{D}_g|} \sum_{i \in \mathcal{D}_g} \ell_i.$$

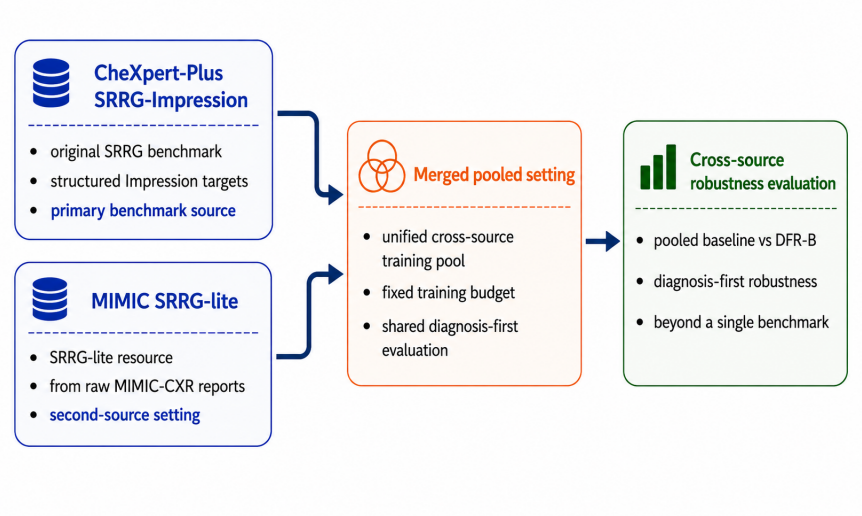


Figure 2: Construction of the pooled cross-source setting. The primary CheXpert-Plus SRRG-Impression benchmark and the MIMIC-based SRRG-lite resource are combined under a fixed training budget and evaluated using the same diagnosis-priority robustness protocol.

The group-aware objective is

$$\mathcal{L}_{\text{group}} = \sum_{g \in \mathcal{G}} q_g \mathcal{L}_g,$$

where q_g controls the influence of each group and places additional emphasis on underperforming groups.

3.5 Diagnosis-First Robust Training

The full DFR objective combines the two mechanisms:

$$\mathcal{L}_{\text{DFR}} = \sum_{g \in \mathcal{G}} q_g \left(\frac{1}{|\mathcal{D}_g|} \sum_{i \in \mathcal{D}_g} w_i^{\text{df}} \ell_i \right).$$

We evaluate four training variants: **baseline**, **df_only**, **group_only**, and **full DFR**. We additionally evaluate two full-DFR configurations, DFR-A and DFR-B, to characterize different robustness operating points. DFR-B is the default configuration unless otherwise stated.

3.6 Cross-source setting

To test whether diagnosis-priority robustness remains visible beyond one benchmark-specific distribution, we evaluate three regimes. The primary setting uses CheXpert-Plus SRRG-Impression. The second-source setting uses an SRRG-lite resource derived from MIMIC-CXR reports. The pooled setting combines the two sources into one training regime while keeping the training budget fixed. The same diagnosis-priority evaluation protocol is used across all settings. Figure 2 illustrates the construction of the pooled cross-source setting.

4 Experiments

4.1 Experimental setup

We evaluate T5-Base and T5-Flan backbones. On the primary benchmark, we compare the baseline objective, diagnosis-priority weighting only, group-aware reweighting only, and full DFR. Multi-seed experiments use seeds $\{0, 1, 42\}$ when available.

Table 1: Main results on the primary SRRG benchmark at $\tau = 2000$. Higher is better except Rare FN Rate.

Backbone	Method	BLEU	ROUGE-L	BERTScore	Leaf F1	Upper F1	Worst F1	G11 F1	Rare Recall	Rare FN
Base	baseline	0.466	0.634	0.940	0.858	0.860	0.781	0.837	0.735	0.265
Base	df_only	0.479	0.643	0.941	0.860	0.861	0.808	0.835	0.733	0.267
Base	group_only	0.479	0.642	0.941	0.860	0.860	0.796	0.840	0.738	0.262
Base	full_dfr(B)	0.488	0.641	0.941	0.858	0.857	0.820	0.837	0.723	0.277
Flan	baseline	0.469	0.634	0.939	0.860	0.852	0.789	0.844	0.740	0.260

Table 2: Multi-seed robustness evaluation for T5-Base at $\tau = 2000$. Values are mean $\pm$ standard deviation. Higher is better except Rare FN Rate.

Method	ROUGE-L	Leaf F1	Worst F1	G11 F1	Rare Recall	Rare FN
baseline	0.634 ± 0.002	0.858 ± 0.001	0.781 ± 0.006	0.837 ± 0.008	0.735 ± 0.009	0.265 ± 0.009
df_only	0.643 ± 0.007	0.860 ± 0.003	0.808 ± 0.014	0.835 ± 0.007	0.733 ± 0.015	0.267 ± 0.015
group_only	0.642 ± 0.007	0.860 ± 0.005	0.796 ± 0.003	0.840 ± 0.010	0.738 ± 0.019	0.262 ± 0.019
full_dfr(B)	0.641 ± 0.012	0.858 ± 0.004	0.820 ± 0.024	0.837 ± 0.003	0.723 ± 0.025	0.277 ± 0.025

We report three categories of metrics. First, we measure aggregate text-generation quality with BLEU, ROUGE-L, and BERTScore. Second, we report SRR-BERT semantic metrics, including Leaf F1 and Upper F1. Third, we report diagnosis-priority robustness using Worst-group F1, Group 11 F1, Rare Recall, and Rare False-Negative Rate. Higher is better for all metrics except Rare False-Negative Rate.

4.2 Main results on the primary benchmark

Table 1 reports the primary results. Aggregate text and semantic metrics remain relatively close across training variants, but diagnosis-priority metrics separate the resulting operating points more clearly.

The full DFR objective improves Worst-group F1 from 0.781 to 0.820 while preserving similar aggregate semantic quality. However, the improvement is not uniform across all diagnosis-priority metrics. The group-only variant obtains the highest Group 11 F1 and Rare Recall, while full DFR places greater emphasis on the worst-performing operative group. This distinction is central to our interpretation: diagnosis-priority training changes the balance among robustness objectives rather than producing a single configuration that dominates every metric.

The T5-Flan baseline further illustrates the same issue. Its aggregate semantic quality is competitive with T5-Base, but its worst-group behavior remains close to the baseline operating point. Aggregate quality therefore does not uniquely determine diagnosis-priority reliability.

4.3 Multi-seed robustness analysis

Table 2 reports multi-seed results for T5-Base. The ranking of robustness profiles is not restricted to a single random seed. Full DFR achieves the strongest mean Worst-group F1, while group-only training obtains the strongest mean Group 11 F1 and Rare Recall.

These results reinforce the operating-point interpretation. The effect of diagnosis-priority optimization is visible most clearly in how performance is redistributed across robustness metrics, not in a uniform increase of every score.

4.4 Operating-point analysis

We next compare two full-DFR configurations. DFR-A places relatively more emphasis on rare-case behavior, while DFR-B emphasizes worst-group performance.Figure 3 visualizes this operating-point trade-off.

The two configurations remain close in aggregate semantic quality while occupying different diagnosis-priority operating points. This result argues against summarizing model quality with a

Table 3: A/B operating-point comparison for full DFR on the primary benchmark.

Method	ROUGE-L	Leaf F1	Worst F1	G11 F1	Rare Recall	Rare FN
DFR-A	0.642	0.860	0.803	0.832	0.730	0.270
DFR-B	0.641	0.858	0.820	0.837	0.723	0.277

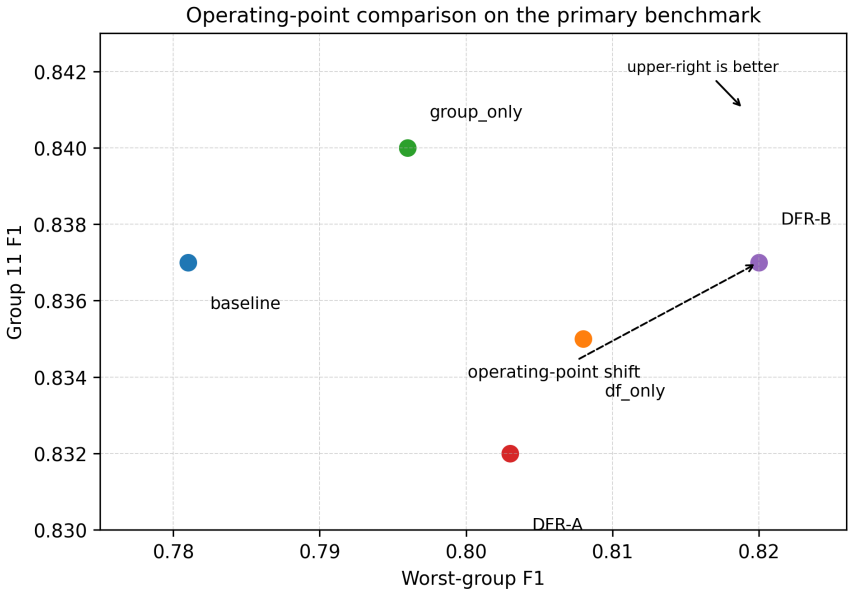


Figure 3: Diagnosis-priority operating points on the primary SRRG benchmark. DFR-A and DFR-B emphasize different robustness dimensions. DFR-B achieves stronger worst-group and Group 11 F1, whereas DFR-A retains higher rare recall. The comparison illustrates that diagnosis-priority robustness is better characterized as an operating-point trade-off than as uniform improvement across metrics.

single aggregate metric. In particular, stronger worst-group performance need not coincide with the strongest rare recall, and vice versa.

4.5 Cross-source and pooled evaluation

We next evaluate whether the same robustness behavior remains observable outside the primary benchmark. Table 4 reports results for the SRRG-lite and pooled settings.

The second-source setting shows that the effect of DFR is not uniform across sources. DFR slightly reduces Worst-group and Group 11 F1 in this setting, while increasing Rare Recall from 0.828 to 0.848 and reducing Rare FN Rate from 0.172 to 0.152. This mixed profile reinforces the need to report multiple robustness dimensions rather than interpreting one metric as a complete measure of reliability.

The pooled setting yields the clearest joint improvement. DFR increases Worst-group F1 from 0.836 to 0.867, increases Rare Recall from 0.733 to 0.757, and reduces Rare FN Rate from 0.267 to 0.243, while ROUGE-L and Leaf F1 remain competitive. We interpret this result as evidence that diagnosis-priority robustness remains meaningful under pooled cross-source data. We do not claim that heterogeneity itself causes the improvement, since that would require a controlled source-mixture experiment.

Table 4: Cross-source robustness results at $\tau = 2000$. Higher is better except Rare FN Rate.

Setting	Method	ROUGE-L	Leaf F1	Worst F1	G11 F1	Rare Recall	Rare FN
Path2	baseline	0.895	0.941	0.925	0.925	0.828	0.172
Path2	full_dfr(B)	0.892	0.941	0.921	0.921	0.848	0.152
Pooled	baseline	0.760	0.889	0.836	0.879	0.733	0.267
Pooled	full_dfr(B)	0.761	0.892	0.867	0.867	0.757	0.243

Table 5: Rare-threshold support analysis under the support-validity rule used in the experiments.

Threshold τ	Group 11 support	Selected
100	3	No
300	6	No
500	12	No
1000	36	No
2000	127	Yes

4.6 Slice statistics and threshold validity

The rarity threshold is chosen using training support rather than validation or test performance. Table 5 summarizes Group 11 support across the predefined threshold grid.

Group distributions are shown in Table 6. Group 11 remains a small but non-degenerate minority slice across settings, while Group 01 is nearly absent. This motivates excluding Group 01 from the main worst-group computation.

5 Discussion

Average quality does not determine diagnosis-priority reliability. The central empirical finding is that models with similar aggregate report quality can occupy meaningfully different robustness operating points. BLEU, ROUGE-L, BERTScore, and semantic F1 vary relatively little across several primary-setting variants, while Worst-group F1, Group 11 F1, and Rare Recall separate them more clearly. Aggregate generation metrics therefore provide an incomplete description of reliability.

Diagnosis-priority robustness is multi-dimensional. The results do not support the claim that one DFR configuration uniformly improves every robustness metric. Instead, diagnosis-priority weighting, group-aware reweighting, and their combination emphasize different failure modes. Group-only training produces the strongest Group 11 F1 and Rare Recall on the primary benchmark, while full DFR produces the strongest Worst-group F1. The A/B comparison shows the same pattern within full DFR. We therefore view robust SRRG as an operating-point problem in which different objectives redistribute performance across clinically motivated reliability criteria.

Cross-source results expose dependence on data composition. The second-source and pooled results show different robustness profiles. In the second-source setting, DFR primarily improves rare recall and false-negative rate, whereas the pooled setting shows joint improvements in worst-group F1, rare recall, and rare false-negative rate. These differences indicate that conclusions drawn from a single benchmark can be incomplete. At the same time, our experiments do not isolate which source-specific factors produce these differences, so we avoid making a causal claim about heterogeneity.

Clinical interpretation. Diagnosis-priority metrics should be interpreted as reliability diagnostics rather than direct measures of patient harm. Rare findings are not necessarily more severe, and a semantic-label error does not uniquely determine a downstream clinical outcome. The purpose of the framework is narrower: to reveal whether clinically motivated minority semantics receive comparable reliability to majority patterns and whether training changes that balance.

Table 6: Diagnosis-priority group distributions across settings and splits.

Setting	Split	Total	G00	G01	G10	G11	G11 (%)
Path1	train	4000	658	0	3154	188	4.7
Path1	val	300	42	0	244	14	4.7
Path1	test	2219	303	0	1789	127	5.7
Path2	train	4000	1342	0	2511	147	3.7
Path2	val	300	90	0	201	9	3.0
Path2	test	2219	408	0	1713	98	4.4
Pooled	train	4000	963	0	2869	168	4.2
Pooled	val	600	142	0	433	25	4.2
Pooled	test	2000	315	0	1594	91	4.5

6 LIMITATIONS

This study has several limitations. First, diagnosis-priority slicing depends on reference-side SRR-BERT semantic labels. Errors or instability in the label extractor can therefore affect both group assignment and semantic evaluation. Second, rarity is defined through training-set support and should not be interpreted as clinical severity. A rare finding can be minor, while a common finding can be clinically urgent. Third, the experiments focus on chest radiography and structured Impression generation. The conclusions may not transfer directly to other imaging modalities, institutions, report formats, or more open-ended clinical generation tasks.

Fourth, DFR is a lightweight reweighting framework. The present study establishes the diagnosis-priority robustness problem and analyzes operating-point behavior, but does not claim that the proposed objective is a universally optimal robust-learning method. More adaptive robust objectives, continuous rarity models, uncertainty-aware weighting, or ontology-guided objectives may yield stronger results. Fifth, cross-source results differ across settings. Although the pooled setting yields the clearest joint improvements, the current experiments do not isolate source heterogeneity as the causal mechanism. Controlled source-mixture experiments would be required for that conclusion.

Finally, automatic semantic metrics cannot fully represent clinical usefulness or harm. Human clinical review and downstream decision studies would be needed to determine whether changes in diagnosis-priority metrics translate into improved clinical outcomes.

7 CONCLUSION

We study Structured Radiology Report Generation beyond average performance by introducing a diagnosis-priority view of semantic-tail robustness. Using reference-side semantic labels, we define abnormality-rarity slices and evaluate both abnormal-and-rare performance and worst-group reliability. We further introduce Diagnosis-First Robust Training, which combines diagnosis-priority weighting with group-aware reweighting.

Across the primary benchmark, multiple random seeds, a second-source setting, and pooled cross-source data, the results show that aggregate generation quality does not uniquely determine diagnosis-priority reliability. Different training objectives occupy different operating points: full DFR produces the strongest worst-group result on the primary benchmark, while other variants favor Group 11 performance or rare recall. Under pooled cross-source training, DFR jointly improves worst-group F1, rare recall, and rare false-negative rate while maintaining competitive aggregate quality.

These findings motivate evaluating structured clinical generators with reliability criteria that explicitly expose low-frequency diagnostic semantics rather than inferring robustness from average text-generation performance alone.

REFERENCES

Pierre Chambon, Jean-Benoit Delbrouck, Tanya Sounack, Samuel Huang, Zhihong Chen, Maya Varma, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text

reports and patient information, 2024. URL https://arxiv.org/abs/2405.19538.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer, 2022. URL https://arxiv.org/abs/2010.16056.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Reis, Mohamed Muneer, Maya Varma, and Curtis Langlotz. Automated structured radiology report generation, 2025. URL https://arxiv.org/abs/2505.24223.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. URL https://arxiv.org/abs/2106.14463.

Alistair E. W. Johnson, Tom J. Pollard, Seth Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://www.nature.com/articles/s41597-019-0322-0.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information, 2021. URL https://arxiv.org/abs/2107.09044.

Qiguang Miao, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, Ruixuan Liu, Tianyi Liu, Kun Xie, and Zhicheng Jiao. Evoke: Elevating chest x-ray report generation via multi-view contrastive learning and patient-specific knowledge, 2025. URL https://arxiv.org/abs/2411.10224.

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.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment, 2022. URL https://arxiv.org/abs/2112.15011.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexrank: A public leaderboard for ai-powered radiology report generation, 2024. URL https://arxiv.org/abs/2411.15122.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. Large model driven radiology report generation with clinical quality reinforcement learning, 2024. URL https://arxiv.org/abs/2403.06728.

A Rare-Threshold Selection and Slice Validity

The diagnosis-first evaluation requires a rarity threshold that produces a meaningful abnormal-and-rare slice. We select this threshold using training-data support rather than validation or test performance.

We evaluate the predefined threshold grid

$$\tau \in \{100, 300, 500, 1000, 2000\}.$$

For each threshold, semantic labels whose training support falls below the threshold are treated as rare. The corresponding abnormality–rarity groups are then constructed from reference-side semantic labels.

Very small thresholds yield nearly degenerate Group 11 subsets, making group-specific F1 unstable and difficult to interpret. The selected threshold $\tau = 2000$ is the first setting used in the main experiments that provides a sufficiently supported abnormal-and-rare slice under the predefined support-validity rule. Validation and test performance are not used for threshold selection.

Table 7: Rare-threshold support analysis.

Threshold τ	Group 11 support	Selected
100	3	No
300	6	No
500	12	No
1000	36	No
2000	127	Yes

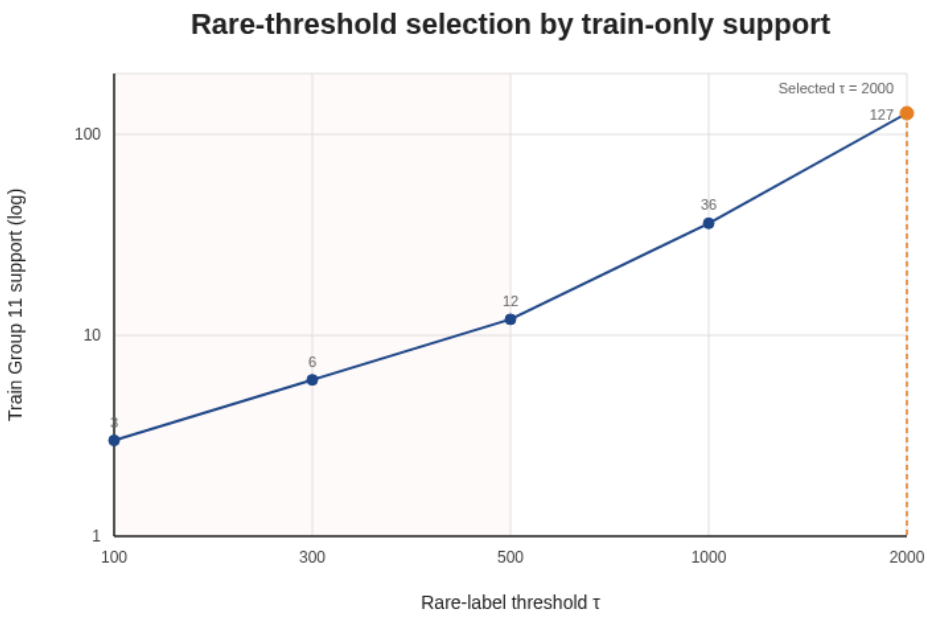


Figure 4: Rare-threshold support across the predefined threshold grid. The selected threshold $\tau = 2000$ yields a non-degenerate abnormal-and-rare slice under the training-support rule.

B Diagnosis-Priority Group Statistics

Table 8 reports the diagnosis-priority group distribution for every setting and split. Group 11 remains a small but non-degenerate minority slice across the primary, second-source, and pooled settings.

Group 01 is absent or nearly absent under the adopted slicing definition. Because a near-empty group would make a minimum-over-groups statistic unstable or undefined, Group 01 is excluded from the principal worst-group calculation. We therefore define

$$\mathrm{WorstF1} = \min_{g \in \{00,10,11\}} \mathrm{F1}_g.$$

Table 8: Group distribution under diagnosis-first slicing across settings and splits.

Setting	Split	Total	Group 00	Group 01	Group 10	Group 11	Group 11 %
Path1	train	4,000	658	0	3,154	188	4.7
Path1	val	300	42	0	244	14	4.7
Path1	test	2,219	303	0	1,789	127	5.7
Path2	train	4,000	1,342	0	2,511	147	3.7
Path2	val	300	90	0	201	9	3.0
Path2	test	2,219	408	0	1,713	98	4.4
Pooled	train	4,000	963	0	2,869	168	4.2
Pooled	val	600	142	0	433	25	4.2
Pooled	test	2,000	315	0	1,594	91	4.5

C DFR Variants and Training Modes

DFR consists of two independently controlled components. Diagnosis-priority weighting is controlled by α, while group-aware reweighting is controlled by η. This produces four training modes:

Table 9: DFR training variants.

Method	Diagnosis-priority weighting	Group reweighting
baseline	No	No
df_only	Yes	No
group_only	No	Yes
full_dfr	Yes	Yes

The baseline corresponds to standard average-risk fine-tuning. The `df_only` variant isolates the contribution of diagnosis-priority weighting. The `group_only` variant isolates group-aware reweighting. The full DFR model combines both mechanisms.

For the principal DFR implementation, the initial configuration uses $\alpha = 0.5$ and $\eta = 0.1$. Two full-DFR operating-point configurations are additionally evaluated to characterize the trade-off among worst-group performance, Group 11 performance, and rare recall.

D Additional Operating-Point Results

Table 10 compares two full-DFR configurations. The configurations produce similar aggregate semantic quality but different diagnosis-priority behavior.

Table 10: A/B trade-off for full DFR on the primary benchmark.

Method	ROUGE-L	Leaf F1	Worst F1	Group 11 F1	Rare Recall	Rare FN
full_dfr(A)	0.642	0.860	0.803	0.832	0.730	0.270
full_dfr(B)	0.641	0.858	0.820	0.837	0.723	0.277

E SRRG-LITE CONSTRUCTION

For the second-source setting, raw MIMIC-CXR reports are converted into standardized structured Impression targets. The restructuring procedure operates on the original Impression content and maps it into a fixed schema beginning with the header `Impression:`, followed by sequentially numbered statements.

The restructuring is constrained to preserve supported clinical content. It does not intentionally introduce diagnoses, findings, procedures, devices, measurements, medications, or clinical history that are absent from the source report. Negation and uncertainty expressions are retained when supported by the original Impression.

Each processed record contains the structured Impression together with the corresponding source identifier and split information. If the original Impression is empty or cannot be converted into the expected structure, the output uses the fallback representation

```
Impression: 1. Not available.
```

The resulting structured targets are subsequently processed by the same reference-side semantic labeling and diagnosis-priority evaluation pipeline used for the primary setting.

F IMPLEMENTATION AND EVALUATION DETAILS

All diagnosis-priority group assignments are computed before training from reference-side semantic labels. Group assignments are never derived from model predictions.

The operative group set is

$$\{00, 10, 11\}.$$

Group 01 is retained in the general slicing definition but excluded from worst-group computation because it is degenerate in the evaluated data.

Multi-seed experiments use the fixed seed set

$$\{0, 1, 42\}.$$

Mean and standard deviation are reported across available runs.

We evaluate three classes of metrics. Text-generation quality is measured using BLEU, ROUGE-L, and BERTScore. Semantic fidelity is measured using SRR-BERT Leaf F1 and Upper F1. Diagnosis-priority robustness is evaluated using Worst-group F1, Group 11 F1, Rare Recall, and Rare False-Negative Rate.

The same diagnosis-priority slicing and evaluation conventions are used across the primary, second-source, and pooled settings.