DIAGNOSTIC LOSS: ASYMMETRIC SUPERVISION FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated radiology report generation requires models to do more than produce fluent clinical narratives: they must explicitly state supported diagnoses while avoiding unsupported ones. However, existing medical vision-language models (MedVLMs) are typically trained with token-level maximum likelihood objectives, which enforce general text agreement between AI-made and human-made radiology reports but do not prioritize diagnostically critical terms. As a result, models can generate reports that appear human-made but miss key and obvious diagnoses, and introduce wrong diagnoses. We identify this as an objective misalignment between conventional language modeling and the asymmetric diagnostic requirements of radiology reporting. To encode these requirements, we introduce **Diagnostic Loss (DLoss)**, a differentiable, architecture-agnostic objective that operates directly on decoder token probabilities. DLoss separately aggregates probability mass over vocabularies associated with diagnoses supported by the reference report and with unsupported alternatives. Using asymmetric pooling across token positions in the report, it encourages supported diagnoses to receive sufficient probability in at least one position while suppressing spurious probability spikes for unsupported diagnoses across all positions during training. The loss can be applied to existing 3D MedVLMs without modifying their vision backbones, language decoders, or inference procedures, and introduces no additional inference-time computation. We train and evaluate DLoss on 50,000 abdominal CT scans across several 3D MedVLM architectures on both in-distribution and out-of-distribution report-generation benchmarks. Across these settings, DLoss improves diagnostic faithfulness while maintaining competitive performance on conventional language-generation metrics, showing the value of asymmetric sequence-level supervision for radiology report generation. The code for DLoss is available in an anonymous repository.

1 INTRODUCTION

Radiology reporting represents one of the most critical language-generation challenges in medicine (Boecking et al., 2022). Furthermore, reporting is essential across multiple medical imaging modalities, from single-frame 2D X-rays (Tiu et al., 2022; Liu et al., 2024b; MH Nguyen et al., 2026; Yang et al., 2026) to 3D computed tomography (CT) volumes spanning hundreds of axial slices (Płotka et al., 2026; Zhou et al., 2026a; Rokuss et al., 2026). The radiologist must collapse this dense visual evidence into concise text that states, in clinically explicit terms, which diagnoses are present or absent in the image (Jain et al., 2021) (here, a diagnosis may refer to a disease or a radiological finding). The diagnostic value of a report lies precisely in this commitment: downstream clinical decisions, such as surveillance, biopsy, surgery, and oncologic treatment, are triggered by whether specific diagnostic terms appear and whether incorrect ones are omitted (Liu et al., 2024a). Automating elements of this process is highly appealing as imaging volumes continue to outpace radiologist capacity, but such automation is viable only if the generated report preserves the diagnostic value of the original report rather than merely mimicking its style and looking like a human-written report (Zhou et al., 2024; Wang et al., 2026; Bassi et al., 2025).

Recently, vision-language models (VLMs) trained on medical image-report pairs (Shui et al., 2025; Cao et al., 2025; Claessens et al., 2026; Zhao et al., 2026; Hamamci et al., 2026a;b; Blankemeier et al., 2026) have brought this goal significantly closer. By coupling visual encoders with large lan-

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.

guage models, modern medical VLMs (MedVLMs) produce fluent, well-structured reports and have been increasingly applied for report generation (Shui et al., 2025; Cao et al., 2025; Kalisch et al., 2025; Hamamci et al., 2026a; Blankemeier et al., 2026; Yan et al., 2026). However, a recurring failure mode persists across architectures: generated reports can look very similar to human-written reports while omitting diagnoses that are present in the image, under-calling their severity, or introducing diagnoses that the image does not support (Delbrouck et al., 2022; Yu et al., 2023; Ostmeier et al., 2024; Cheng & Subramanian, 2026; Bassi et al., 2025). Radiologist audits of generated chest radiographs score these as clinically significant errors, false diagnoses and omitted diagnoses, precisely when they would alter clinical workflow (Yu et al., 2023; Ostmeier et al., 2024). In radiology and clinical oncology, missing or weakening an actionable diagnosis is more consequential than a stylistic error, because subsequent staging and treatment depend on whether the disease is named and on whether unsupported differentials are introduced (Bruno et al., 2015; Bruno, 2017).

This mismatch between generating reports that look human-written and achieving diagnostic correctness is not explained by modeling capacity alone; it also reflects a misalignment in the training objective (Miura et al., 2021; Delbrouck et al., 2022; Cheng & Subramanian, 2026). Standard maximum-likelihood training optimizes token-level cross-entropy against a ground-truth report (Jing et al., 2018; Delbrouck et al., 2022). This objective treats diagnostic and non-diagnostic tokens through the same token-level loss: frequent narrative tokens dominate the gradient, rare diagnostic terms are under-weighted, and there is no explicit sequence-level supervision encouraging supported diagnostic terms to receive sufficient probability at some token position in the report while suppressing probability peaks for unsupported terms across all positions (Welleck et al., 2020; Wu et al., 2023; Liu et al., 2026). There is, therefore, little training-time pressure to satisfy the radiologist's two-sided diagnostic obligation—supported diagnoses must be named, and unsupported diagnoses must not be introduced (Hartung et al., 2020; Yu et al., 2023; Ostmeier et al., 2024). Consequently, even strong MedVLMs can produce diagnostically incomplete or misleading reports despite strong scores on conventional language metrics (Papineni et al., 2002; Lin, 2004; Miura et al., 2021; Delbrouck et al., 2022; Yu et al., 2023).

To address this misalignment in the training objective, we introduce Diagnostic Loss (DLoss), an auxiliary objective for radiology report generation that directly targets this two-sided diagnostic requirement: enforcing correct diagnoses and penalizing incorrect ones. DLoss is fully differentiable, operates directly on the decoder's token distribution, and is architecture-agnostic. It can be integrated on top of 3D autoregressive MedVLMs that expose decoder logits without altering the underlying backbone, tokenizer, or report-level supervision. Concretely, it utilizes asymmetric pooling of token probabilities over per-sample diagnostic vocabularies: probability mass on supported diagnostic targets is aggregated across token positions in the report to reward expressing the correct diagnosis somewhere, while probability mass on unsupported diagnostic targets is pooled with a tempered soft maximum to penalize any spurious spikes across these positions during training. This formulation encodes the asymmetric requirements of supported and unsupported diagnostic targets. Our main contributions are as follows:

1. **Asymmetric diagnostic supervision.** We introduce DLoss, which encodes different sequence-level constraints for supported and unsupported diagnostic concepts through asymmetric pooling of decoder token probabilities across token positions in the report. Supported concepts are encouraged to receive sufficient probability at least once, while unsupported concepts are penalized for probability spikes across these positions during training.
2. **Tokenizer-agnostic concept supervision.** We represent diagnostic targets at the concept level and map their associated words and phrases (e.g., *gallstones* and *cholelithiasis*) to the token vocabulary of each decoder using its existing tokenizer. This provides token-level diagnostic supervision while allowing the same objective to operate across different tokenizers and MedVLM architectures without introducing a task-specific prediction head or modifying the inference procedure.
3. **Cross-model and out-of-distribution evaluation.** We train and evaluate DLoss on 50,000 abdominal CT scans across several 3D MedVLMs, decoder and tokenizer architectures, and both in- and out-of-distribution report-generation benchmarks. In-distribution experiments are conducted on abdominal CT scans (Merlin), while out-of-distribution performance is evaluated on a private, in-house abdominal CT dataset.

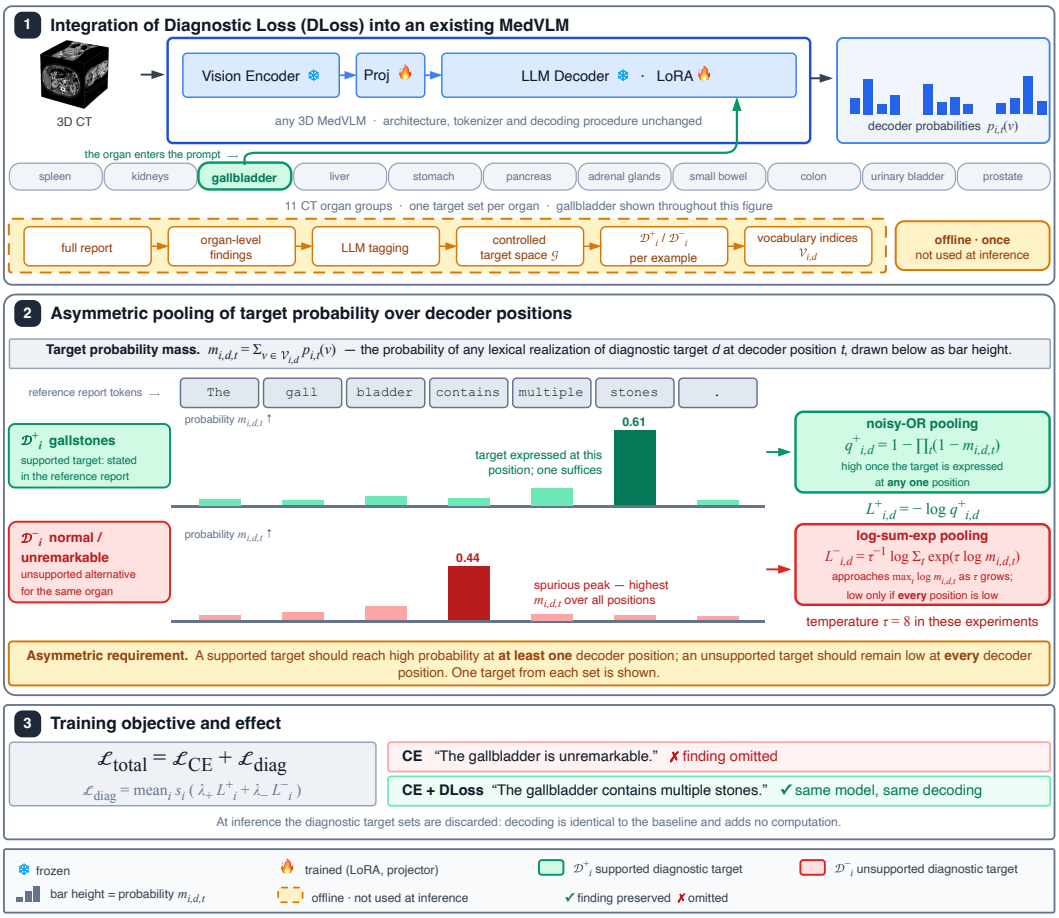


Figure 1: **Overview of DLoss and its workflow.** DLoss supervises how probability is distributed over diagnostic targets alongside standard token prediction. **(1) Integration.** DLoss operates directly on decoder probabilities without modifying the architecture or decoding procedure (only the LoRA parameters and the projector are trained). Reports are decomposed into organ-level findings and mapped offline, via LLM tagging, to supported ($\mathcal{D}_i^+$) and unsupported ($\mathcal{D}_i^-$) diagnostic targets in a controlled target space $\mathcal{G}$. **(2) Asymmetric pooling.** Target probability masses $m_{i,d,t}$ across valid decoder positions are aggregated using distinct operators: noisy-OR for supported targets (saturates when present at any position) and temperature-controlled log-sum-exp for unsupported targets (ensures probability stays low everywhere). **(3) Objective.** The supported and unsupported target losses are added to cross-entropy with independent coefficients λ_+ and λ_-, respectively, and scaled by sample weight s_i. The diagnostic target sets are discarded at inference, so DLoss adds no inference-time computation.

Across six 3D MedVLMs (Merlin, fVLM, BTB3D, HLIP, RadFM, and SPECTRE), adding DLoss yields mean gains of 4.05 GREEN points and 3.82 RadGraph-F1 points on the Merlin test set, and 4.28 and 2.96 points, respectively, on the external test set (see Table 1). Mean improvements across BLEU, ROUGE-L, METEOR, and CIDEr also show consistent positive gains across both sets.

2 Related Work

MedVLMs for Radiology Report Generation. Radiology report generation has evolved from specialized image-to-text architectures toward more general medical vision–language models. Early 2D approaches used CNN–RNN or hierarchical recurrent architectures with attention (Wang et al., 2018a; Jing et al., 2018), followed by Transformer-based generators such as R2Gen (Chen et al.,

2020). More recent approaches leverage pretrained vision–language representations, such as BiomedCLIP (Zhang et al., 2025), and generative VLMs or large language models, including LLM-CXR (Lee et al., 2024), CMCRL (Chen et al., 2025), MLRG (Liu et al., 2025b), and AOR (Li et al., 2026a). For volumetric CT, recent work includes fVLM (Shui et al., 2025), RadFM (Wu et al., 2025), HLIP (Zhao et al., 2026), BTB3D (Hamamci et al., 2026a), SPECTRE (Claessens et al., 2026), and Merlin (Blankemeier et al., 2026). DLoss is complementary to these architectural and representation-learning advances, modifying supervision over the decoder distribution rather than the underlying vision or language architecture.

Clinical Supervision and Training Objectives. Clinical fidelity is not fully captured by lexical similarity, motivating measures such as RadGraph-based scores (Jain et al., 2021; Delbrouck et al., 2022) and GREEN (Ostmeier et al., 2024). RadGraph-F1 and GREEN measure clinical relevance by extracting structured entity-relation knowledge graphs and leveraging domain-tuned models to explicitly penalize diagnostic errors like false positives, false negatives, or anatomical misallocations that traditional lexical metrics ignore. Clinical information has also been incorporated into training through reinforcement learning, differentiable disease supervision, and semantic rewards (Liu et al., 2019; Lovelace & Mortazavi, 2020; Delbrouck et al., 2022), token-level reweighting or unlikelihood (Wu et al., 2023; Welleck et al., 2020), disease- or anatomy-aware generation (Jin et al., 2024; Cheng & Subramanian, 2026), and preference or reinforcement learning (Liu et al., 2025a; Zhou et al., 2026b).

Among approaches that modify the training objective, SAT-RRG uses an LLM to locate semantic errors and adjust training at the affected token positions (Liu et al., 2026). Unlikelihood training lowers the probability of selected negative candidates, including in sequence-level variants (Welleck et al., 2020). DLoss instead defines supported and unsupported diagnostic targets from each reference report and pools the decoder probabilities for each target across positions. Its two pooling operations encourage probability for a supported target without specifying where it must occur, while limiting probability spikes for unsupported targets throughout the training sequence. The pooling operations draw on multiple-instance learning (Maron & Lozano-Pérez, 1997; Wang et al., 2018b; Pinheiro & Collobert, 2015).

3 Methodology

3.1 Preliminaries

DLoss augments standard autoregressive report generation with diagnostic supervision applied directly to the language decoder. The method does not modify the visual encoder, language decoder, tokenizer, or inference procedure. Instead, it operates on the decoder probabilities already computed during report generation training.

Autoregressive Report Generation. Consider a minibatch of B radiology report generation examples. Each example i is represented as (x_i, y_i), where x_i denotes the conditioning input, including the image-derived representation and textual prompt, and $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ denotes the reference report tokenized over a vocabulary of size V. At output position t, the language decoder conditions on x_i and the reference prefix $y_{i,<t}$ to produce a logit vector $\boldsymbol{\ell}_{i,t} \in \mathbb{R}^V$. We denote the probability assigned to vocabulary item v by:

$$p_{i,t}(v) = \mathrm{softmax}(\boldsymbol{\ell}_{i,t})_v. \tag{1}$$

Let $\mathcal{P}_i$ denote the set of valid report-token positions for example i, excluding prompt, padding, and image positions. The baseline training objective is the token-level cross-entropy loss over the minibatch:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{\sum_{i=1}^{B} |\mathcal{P}_i|} \sum_{i=1}^{B} \sum_{t \in \mathcal{P}_i} \log p_{i,t}(y_{i,t}). \tag{2}$$

Cross-entropy provides local supervision for each reference token, but does not explicitly distinguish diagnostically relevant concepts from the remaining report text. DLoss complements this

objective with supervision over diagnostic targets that should receive probability during decoding and diagnostic targets that should remain unlikely.

Anatomy-level supervision for 3D CT. For the 3D CT setting, we define diagnostic supervision at the anatomical level, following prior anatomy-aware CT vision–language approaches (Shui et al., 2025; Cao et al., 2025). Each study–report pair is decomposed into anatomy-level image–text pairs:

$$(\mathbf{X}_n^I, \mathbf{X}_n^T) \longmapsto \left\{(\mathbf{X}_{n,c}^I, \mathbf{X}_{n,c}^T)\right\}_{c\in\mathcal{C}_n}, \tag{3}$$

where $\mathbf{X}_{n,c}^I$ and $\mathbf{X}_{n,c}^T$ denote the image region and corresponding findings for anatomical organs c. Anatomical structures are segmented using TotalSegmentator (Wasserthal et al., 2023), while reports are parsed into organ-level findings using Qwen3-30B-A3B. This decomposition also sets the scope of the loss: supported and unsupported targets are assigned within each organ-level pair and evaluated only over its report-token positions. Applying all organ targets to one full-report sequence could give opposing signals for the same word (e.g., in a report describing a liver cyst and normal kidneys, *cyst* would be encouraged for the liver and penalized for the kidneys).

Diagnostic target construction. Let $\mathcal{G}$ denote a controlled set of diagnostic targets. For each training example i, we construct two subsets: $\mathcal{D}_i^+ \subseteq \mathcal{G}$ contains diagnostic targets supported by the reference report, whereas $\mathcal{D}_i^- \subseteq \mathcal{G}$ contains unsupported diagnostic targets. In the 3D setting, these sets are constructed within the anatomical unit defined above. Reference-report expressions are mapped to concepts in $\mathcal{G}$: supported concepts populate $\mathcal{D}_i^+$ and unsupported alternatives populate $\mathcal{D}_i^-$. Explicitly negated concepts are excluded from both sets to avoid conflicting lexical supervision.

A diagnostic target represents a clinical concept rather than a single surface form. Each target $d \in \mathcal{D}_i^+ \cup \mathcal{D}_i^-$ is therefore associated with a tokenizer-specific set of vocabulary indices $\mathcal{V}_{i,d} \subseteq \{1, \ldots, V\}$ corresponding to lexical realizations of that concept. For example, lexical expressions such as *cholelithiasis*, *gallstones*, and *stone* can contribute to the same diagnostic target. We construct these lexical sets offline, starting from each normalized target name and adding recurrent words and short phrases from organ-specific training reports assigned to that target through LLM-based semantic tagging and synonym consolidation. We filter non-diagnostic expressions; the resulting sets are grounded in the training corpus and do not cover every possible paraphrase. These lexical realizations are tokenized separately for the tokenizer used by each MedVLM, yielding the corresponding sets $\mathcal{V}_{i,d}$. Appendix C shows this target construction with an organ-specific example.

Importantly, DLoss only requires the decoder logits and the resulting sets $\mathcal{D}_i^+$, $\mathcal{D}_i^-$, and $\mathcal{V}_{i,d}$. Consequently, the objective is independent of the architecture used to obtain the visual representation and applies to any MedVLM that exposes decoder logits. Since lexical realizations are represented through tokenizer vocabulary indices, the formulation provides token-level diagnostic supervision rather than exact phrase-level matching.

3.2 Diagnostic Loss

Radiology reporting imposes asymmetric requirements on supported and unsupported diagnostic targets. During training, a supported diagnostic target should receive sufficient probability at at least one valid decoder position, whereas an unsupported diagnostic target should remain unlikely across all valid decoder positions. DLoss encodes these two requirements using different pooling operators across decoder positions.

Diagnostic target probability mass. For diagnostic target d, we first aggregate the probability mass assigned to its tokenizer-specific vocabulary indices at output position t:

$$m_{i,d,t} = \sum_{v\in\mathcal{V}_{i,d}} p_{i,t}(v). \tag{4}$$

The quantity $m_{i,d,t}$ therefore represents the total probability assigned to lexical evidence associated with diagnostic target d at a particular decoder position.

Supported-Target Pooling. For a supported target $d \in \mathcal{D}_i^+$, evidence at any valid decoder position can support the diagnosis. We aggregate the target probabilities across positions using noisy-OR pooling:

$$q^+_{i,d} = 1 - \prod_{t \in \mathcal{P}_i} (1 - m_{i,d,t}) \,. \tag{5}$$

The corresponding supported-target loss is:

$$L^+_{i,d} = -\log q^+_{i,d}. \tag{6}$$

The pooled probability increases when diagnostic evidence receives probability at one or more valid decoder positions. Minimizing $L^+_{i,d}$ therefore encourages the decoder to assign sufficient probability to a lexical realization of the supported target without prescribing a particular output position.

Unsupported-Target Pooling. For an unsupported target $d \in \mathcal{D}^-_i$, a high probability at even a single position can introduce diagnostically inconsistent content. We therefore aggregate target probabilities using a temperature-controlled log-sum-exp over their logarithms:

$$L^-_{i,d} = \tau^{-1} \log \left(\sum_{t \in \mathcal{P}_i} \exp\left(\tau \log m_{i,d,t}\right) \right), \tag{7}$$

where $\tau > 0$ controls the sharpness of the aggregation. As τ increases, the pooling increasingly concentrates on the output position with the largest unsupported target probability. Minimizing $L^-_{i,d}$ therefore suppresses spurious probability peaks for unsupported diagnostic targets across valid decoder positions. For numerical stability, arguments to logarithms are clamped away from zero in the implementation; this stabilization is omitted from the equations for readability.

Algorithm 1 DLoss computation during decoder training.

Require: Decoder logits $\boldsymbol{\ell}_{i,t}$ and valid positions $\mathcal{P}_i$
Require: Diagnostic target sets $\mathcal{D}^+_i$ and $\mathcal{D}^-_i$
Require: Vocabulary-index sets $\mathcal{V}_{i,d}$
Require: Target weights $w^+_{i,d}, w^-_{i,d}$ and sample weights s_i
Require: Hyperparameters λ_+, λ_-, and τ
Require: Cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$
Ensure: Total training loss $\mathcal{L}_{\mathrm{total}}$

$\mathcal{I}_{\mathrm{diag}} \leftarrow \{i : \mathcal{D}^+_i \cup \mathcal{D}^-_i \neq \varnothing\}$
for $i = 1, \ldots, B$ **do**
  for $t \in \mathcal{P}_i$ **do**
    $p_{i,t}(v) \leftarrow \mathrm{softmax}(\boldsymbol{\ell}_{i,t})_v$
  end for
  for $d \in \mathcal{D}^+_i \cup \mathcal{D}^-_i$ **do**
    for $t \in \mathcal{P}_i$ **do**
      $m_{i,d,t} \leftarrow \sum_{v \in \mathcal{V}_{i,d}} p_{i,t}(v)$
    end for
  end for
  for $d \in \mathcal{D}^+_i$ **do**
    $q^+_{i,d} \leftarrow 1 - \prod_{t \in \mathcal{P}_i} (1 - m_{i,d,t})$
    $L^+_{i,d} \leftarrow -\log q^+_{i,d}$
  end for
  for $d \in \mathcal{D}^-_i$ **do**
    $L^-_{i,d} \leftarrow \tau^{-1} \log\Big(\sum_{t \in \mathcal{P}_i} \exp(\tau \log m_{i,d,t})\Big)$
  end for
  Compute W^+_i and W^-_i from $w^+_{i,d}$ and $w^-_{i,d}$
  Compute L^+_i and L^-_i as weighted means
  $\ell^{\mathrm{diag}}_i \leftarrow s_i \left(\lambda_+ L^+_i + \lambda_- L^-_i\right)$
end for
if $|\mathcal{I}_{\mathrm{diag}}| > 0$ **then**
  $\mathcal{L}_{\mathrm{diag}} \leftarrow \frac{1}{|\mathcal{I}_{\mathrm{diag}}|} \sum_{i \in \mathcal{I}_{\mathrm{diag}}} \ell^{\mathrm{diag}}_i$
else
  $\mathcal{L}_{\mathrm{diag}} \leftarrow 0$
end if
$\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{diag}}$
return $\mathcal{L}_{\mathrm{total}}$

Overall Training Objective. Let $w^+_{i,d}$ and $w^-_{i,d}$ denote non-negative target-level weights, set to one in all reported experiments, and define $W^+_i = \sum_{d \in \mathcal{D}^+_i} w^+_{i,d}$ and $W^-_i = \sum_{d \in \mathcal{D}^-_i} w^-_{i,d}$. The supported and unsupported diagnostic components for example i are computed as weighted means:

$$L^\pm_i = \begin{cases} \dfrac{\sum_{d \in \mathcal{D}^\pm_i} w^\pm_{i,d} L^\pm_{i,d}}{W^\pm_i}, & W^\pm_i > 0, \\ 0, & W^\pm_i = 0. \end{cases} \tag{8}$$

The sample-level diagnostic loss is then:

$$\ell^{\mathrm{diag}}_i = s_i \left(\lambda_+ L^+_i + \lambda_- L^-_i\right), \tag{9}$$

where $s_i \geq 0$ denotes the sample weight, and the non-negative coefficients λ_+ and λ_- independently control the contributions of supported and unsupported diagnostic targets. Let $\mathcal{I}_{\mathrm{diag}} = \{i \in$

$\{1, \ldots, B\} : \mathcal{D}_i^+ \cup \mathcal{D}_i^- \neq \varnothing\}$ denote the set of training examples containing at least one active diagnostic target. The batch-level DLoss is defined as:

$$\mathcal{L}_{\text{diag}} = \begin{cases} \dfrac{1}{|\mathcal{I}_{\text{diag}}|} \displaystyle\sum_{i \in \mathcal{I}_{\text{diag}}} \ell_i^{\text{diag}}, & |\mathcal{I}_{\text{diag}}| > 0, \\ 0, & |\mathcal{I}_{\text{diag}}| = 0. \end{cases} \tag{10}$$

Finally, DLoss is optimized jointly with the standard language-modeling objective: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{diag}}$, where $\mathcal{L}_{\text{diag}}$ contains the weighted supported- and unsupported-target components defined above. Algorithm 1 summarizes the computation of the training objective. DLoss is used only during training. At inference, the diagnostic target sets are discarded and each MedVLM follows the same decoding procedure as its corresponding baseline. The proposed objective therefore introduces no additional inference-time computation and requires no modification of the underlying vision or language architecture.

4 Experiments

4.1 Exprimental setup

Datasets. For training and evaluation, we employ three distinct 3D abdominal CT datasets paired with their corresponding radiology reports. Specifically, models are trained on CancerVerse (Li et al., 2026b) (24,168 scans from 13,863 patients) and the training split of the publicly available Merlin (Blankemeier et al., 2026; Bassi et al., 2026) dataset (20,369 scans). Together, Merlin and CancerVerse represent the largest public abdominal CT datasets, containing over 20 million image slices. Evaluation is conducted on Merlin's official test set (5,125 scans) for in-distribution assessment, alongside an external, in-house dataset (2,706 scans from 1,948 patients) for out-of-distribution evaluation. Comprehensive details for all datasets are provided in Appendix A.

Implementation details. Pooled training CT volumes are provided as a converted 1.5-mm isotropic view. Each backbone then applies model-specific intensity normalization, geometric preprocessing, and feature extraction based on its original implementation. Scans from Merlin's official test split are supplied at native spacing and processed by the corresponding backbone-specific pipeline. Within each backbone, the CE and DLoss runs use identical CT-derived inputs. Training uses up to four NVIDIA GH200 GPUs, each equipped with 96 GB of VRAM. The implementation details for all experiments can be found in Appendix E.1.

Evaluation metrics. We use GREEN (Ostmeier et al., 2024), BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Lavie & Denkowski, 2009), CIDEr (Vedantam et al., 2015), and RadGraph-F1 (Jain et al., 2021) to evaluate the performance of report generation. While metrics like BLEU, ROUGE-L, METEOR, and CIDEr measure surface-level n-gram overlap, RadGraph-F1 and GREEN evaluate diagnostic accuracy. RadGraph-F1 extracts anatomical entities, findings, and their relationships into knowledge graphs to verify factual consistency. GREEN uses a domain-adapted language model to explicitly identify and weight clinically critical errors, such as false positives, false negatives, or anatomical mislocations, aligning its score directly with radiologist judgment rather than phrasing. A detailed description can be found in Appendix E.2.

Enhancing state-of-the-art baselines. To demonstrate the broad utility and fair contribution of our proposed DLoss, we integrate it across a diverse set of leading MedVLMs, including Merlin (Blankemeier et al., 2026), SPECTRE (Claessens et al., 2026), RadFM (Wu et al., 2025), fVLM (Shui et al., 2025), BTB3D (Hamamci et al., 2026a), and HLIP (Zhao et al., 2026). Crucially, we evaluate each model fine-tuned with vs. without DLoss under identical experimental conditions, directly isolating the performance gains attributable to our diagnostic objective from standard supervised fine-tuning.

4.2 Experimental results

Quantitative results. Table 1 shows that adding DLoss improves report generation across six 3D MedVLMs on both the Merlin and external test sets. Averaged over backbones, GREEN and F1-RG

Table 1: **Quantitative evaluation of 3D radiology report generation on internal and external benchmarks.** We report language generation and clinical accuracy metrics on the Merlin (**M**) test set and an external (**E**) in-house dataset. F1-RG denotes the complete RadGraph score. Rows labeled **+ DLoss** reflect signed changes relative to the baseline model after training with DLoss. Gains are shown in green and drops in red. $\bar{\Delta}$ represents the mean change over the corresponding baseline.

Method	BLEU ↑		ROUGE-L ↑		METEOR ↑		CIDEr ↑		GREEN ↑		F1-RG ↑	
	M	E	M	E	M	E	M	E	M	E	M	E
Merlin	20.65	16.63	44.21	34.30	18.45	15.21	13.98	10.89	53.82	45.01	35.01	26.36
+ DLoss	+0.01	+0.09	+0.05	+0.26	+0.01	+0.07	+0.18	+0.50	+0.83	+1.24	−0.01	+0.35
fVLM	39.58	18.61	50.02	34.06	26.32	15.75	36.25	18.15	53.91	46.78	46.53	29.32
+ DLoss	+7.04	+2.59	+6.70	+2.39	+4.04	+1.26	+6.56	+1.34	+8.99	+3.06	+6.56	+1.65
BTB3D	15.96	16.38	32.21	33.75	17.07	16.14	17.39	19.50	53.38	55.76	25.82	27.95
+ DLoss	+8.62	+6.74	+7.47	+4.39	+3.99	+2.42	+10.52	+4.32	+2.05	+2.06	+9.18	+4.95
HLIP	40.18	21.54	52.16	40.70	26.82	18.07	37.33	21.73	56.15	49.54	48.02	34.22
+ DLoss	+2.44	+2.98	+2.02	+3.47	+1.51	+1.61	+1.74	+2.40	+4.09	+4.43	+2.18	+3.38
RadFM	49.93	27.80	60.55	45.09	32.84	21.06	48.04	27.41	64.80	62.71	57.91	39.74
+ DLoss	+0.06	+0.21	+0.22	+0.48	+0.06	+0.18	+0.11	+0.25	+0.50	+0.86	+0.23	+0.41
SPECTRE	31.88	14.20	42.54	28.14	22.89	12.64	32.30	13.58	45.35	41.34	41.95	26.53
+ DLoss	+5.11	+10.86	+5.41	+10.95	+2.79	+6.59	+3.01	+10.12	+7.84	+14.03	+4.79	+7.03
$\bar{\Delta}$	**+3.88**	**+3.91**	**+3.65**	**+3.66**	**+2.07**	**+2.02**	**+3.69**	**+3.16**	**+4.05**	**+4.28**	**+3.82**	**+2.96**

increase by 4.05 and 3.82 points on Merlin, and by 4.28 and 2.96 points externally. Mean gains are also positive for BLEU, ROUGE-L, METEOR, and CIDEr on both sets. Overall, 71 of the 72 model–metric–test comparisons improve, including every comparison on the external set. The sole exception is a 0.01-point decrease in F1-RG for the Merlin backbone. A complementary nine-seed analysis of training variability and bootstrap uncertainty is provided in Appendix F.

Computational efficiency. Table 2 compares CE training with CE + DLoss across four 3D Med-VLMs. The largest measured mean increase in step time is 4.4% (BTB3D); peak allocated and reserved memory increase by at most 1.6% and 2.9%, respectively. Since DLoss is used only during training, it adds no inference-time computation.

Qualitative results. Figure 2 compares organ-level report excerpts across four CT studies. The baseline model generates false positive findings that are absent in the ground truth references. Conversely, it can also produce false negatives by missing key clinical status updates. Incorporating DLoss addresses both issues by effectively mitigating false positives and false negatives, enabling accurate detection, classification, and reporting of diagnostic findings aligned with the reference standard. These examples show how DLoss enhances overall diagnostic faithfulness by eliminating unsupported findings while consistently retaining true positive and true negative findings.

4.3 Ablation Studies

Table 2: **Measured training-step cost of CE and CE plus DLoss.** Profiling used matched batches of eight in BF16 on a single NVIDIA GH200 GPU. Step times average 12 fresh-process runs per condition across six counterbalanced blocks; time Δ is the mean paired relative change across blocks. Peak allocated and reserved memory average three separate runs per condition, with Δ reporting relative changes between condition means. Allocated memory measures live tensors, while reserved memory includes allocator-held capacity. The optimized DLoss implementations used for profiling were validated against the training objective to ensure identical loss values and gradients.

Method	Step time (ms)			Peak allocated (GiB)			Peak reserved (GiB)		
	CE	+ DLoss	Δ (%)	CE	+ DLoss	Δ (%)	CE	+ DLoss	Δ (%)
fVLM	384.54	391.39	+1.82	3.913	3.976	+1.62	4.422	4.441	+0.44
BTB3D	434.88	453.96	+4.41	56.770	56.775	+0.01	58.375	60.055	+2.88
HLIP	414.97	413.24	−0.36	3.420	3.420	0.00	3.793	3.852	+1.54
RadFM	687.49	690.17	+0.43	27.204	27.411	+0.76	27.842	28.018	+0.63

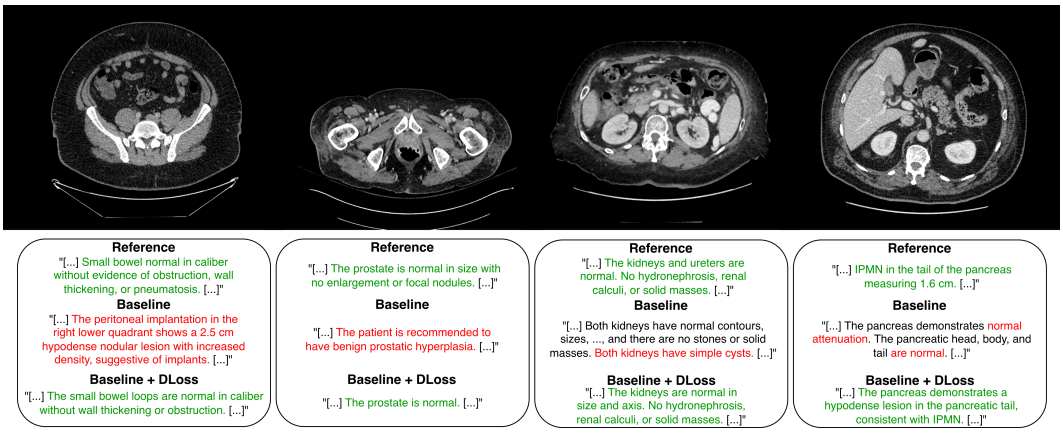


Figure 2: **Qualitative comparison of organ-level CT report generation.** Each column shows a representative CT slice and excerpts from the reference, baseline, and baseline + DLoss findings for one anatomical organ. Green text indicates content consistent with the reference; red text marks baseline statements not supported by the reference.

Table 3 evaluates component and pooling variants of DLoss on fVLM. These ablations use random seed 13, running on the full Merlin test set against their matching CE baseline. At $(\lambda_+, \lambda_-) = (0.001, 0.00025)$, the full DLoss objective improves BLEU-4, GREEN, and F1-RG over the matched CE baseline. Using only the supported (S-only) or unsupported (U-only) target loss degrades performance across all metrics, supporting the complementary contributions of both components in this setting. Replacing supported-target noisy-OR with maximum (S-max) or mean (S-mean), or unsupported-target log-sum-exp with mean (U-mean), also lowers all three scores. U-mean produces the largest declines. Further details are in Appendix G.

Table 3: **Component and pooling ablations of DLoss on fVLM.** S/U denote supported/unsupported targets. "Only" retains CE and the indicated DLoss component; "max" and "mean" replace that target's pooling operator. The CE row gives absolute scores; other rows give relative percentage changes from CE.

Config	BLEU-4 (%)	GREEN (%)	F1-RG (%)
#1 CE	49.55	65.96	55.55
#2 S-only	−1.88	−1.11	−0.74
#3 U-only	−2.75	−2.62	−1.87
#4 S-max	−16.11	−14.76	−11.10
#5 S-mean	−2.27	−1.28	−1.91
#6 U-mean	−22.85	−25.17	−16.37
#7 **DLoss**	**+0.73**	**+0.47**	**+1.54**

5 Conclusions

In this work, we present DLoss, a plug-and-play 3D model-agnostic loss function based on the asymmetric pooling of token probabilities across decoder positions for radiology report generation. Extensive experiments across both in-distribution and out-of-distribution benchmarks show that integrating DLoss consistently enhances diagnostic faithfulness while maintaining competitive language-generation performance. Overall, this lightweight objective improves the performance of six state-of-the-art report generation methods across six evaluation metrics.

Limitations. The controlled diagnostic target space is derived from training reports, so the auxiliary loss directly supervises only the concepts and expressions represented there. Our experiments span six backbones and two test sets but remain within abdominal CT, and the reported scores rely on automatic metrics. Testing DLoss on 2D X-ray reports and comparing its outputs with radiologist judgments would clarify how broadly the observed gains translate.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refining manuscript wording and drafting code. The authors designed the study, conducted the experiments, and determined the interpretation of the results. References were selected by the authors, and their bibliographic entries were copied manually from source records. The authors reviewed AI-assisted material before incorporating it and take full responsibility for the methods, results, and content of this paper.

ETHICS STATEMENT

In this work, we used a private dataset for external evaluation. This study was approved by our Institutional Review Board (IRB); however, to maintain double-blind review in accordance with ICLR policy, the IRB approval number is omitted here and will be added upon acceptance.

REPRODUCIBILITY STATEMENT

The diagnostic objective and training algorithm are specified in Section 3.2. The appendix documents dataset composition and splits (Appendix A), report preprocessing and target construction (Appendices B and C), model and training details (Appendix E.1), evaluation metrics (Appendix E.2), and the paired-seed analysis (Appendix F). An anonymized repository will provide the implementation at https://anonymous.4open.science/r/diagnostic-loss-C5F7/README.md. The in-house test cohort is not publicly available.

REFERENCES

Pedro R A S Bassi, Wenxuan Li, Szymon Płotka, Ruby Honjol, Jakub Przado, Xinze Zhou, Kang Wang, Yang Yang, Malte Jensen, Akshay S Chaudhari, Curtis P Langlotz, Alan L Yuille, and Zongwei Zhou. Merlin plus: A large-scale, multi-cancer, image-mask-report dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 194–204. Springer, 2026.

Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, et al. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23720–23730, 2025.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652 (8112):1318–1328, 2026.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Michael A. Bruno. Communicating uncertainty in the radiology report. *American Journal of Roentgenology*, 209(5):1006–1008, 2017.

Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics*, 35(6):1668–1676, 2015.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 2025.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1439–1449, 2020.

Sheng Cheng and Devika Subramanian. Rethinking radiology report generation: From narrative flow to topic-guided findings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Cris Claessens, Christiaan Viviers, Giacomo D'Amicantonio, Egor Bondarev, and Fons van der Sommen. Scaling self-supervised and cross-modal pretraining for volumetric ct transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13636–13647, 2026.

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

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *Advances in Neural Information Processing Systems*, 38:135074–135102, 2026a.

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*, pp. 1–19, 2026b.

Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. How to create a great radiology report. *Radiographics*, 40(6):1658–1670, 2020.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

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

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 2577–2586, 2018.

Hamza Kalisch, Fabian Hörst, Jens Kleesiek, Ken Herrmann, and Constantin Seibold. Ct-graph: Hierarchical graph attention network for anatomy-guided ct report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6834–6843, 2025.

Alon Lavie and Michael J Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115, 2009.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 29745–29765, 2024.

Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Shang Gao, et al. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *Advances in Neural Information Processing Systems*, 38:19331–19371, 2026a.

Wenxuan Li, Pedro R A S Bassi, Xinze Zhou, Jakob Wasserthal, Ibrahim E Hamamci, Sezgin Er, Bjoern Menze, Gulhan E Akan, Szymon Płotka, Jakub Przado, Qi Chen, Yucheng Tang, Daguang Xu, Arkadiusz Sitek, Kang Wang, Yang Yang, Alan L Yuille, and Zongwei Zhou. CancerVerse: A fully open longitudinal and multimodal dataset for multicancer screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 77–87. Springer, 2026b.

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

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18635–18643, 2024a.

Che Liu, Cheng Ouyang, Sibo Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. G2d: From global to dense radiography representation learning via vision-language pre-training. *Advances in Neural Information Processing Systems*, 37:14751–14773, 2024b.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine learning for healthcare conference*, pp. 249–269. PMLR, 2019.

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

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

Yunyi Liu, Yingshu Li, Tong Chen, Lingqiao Liu, Lei Wang, and Luping Zhou. Sat-rrg: Llm-guided self-adaptive training for radiology report generation with token-level push-pull optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35363–35372, 2026.

Justin Lovelace and Bobak Mortazavi. Learning to generate clinically coherent chest x-ray reports. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 1235–1243, 2020.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

Duy MH Nguyen, Nghiem Diep, Trung Nguyen, Hoang-Bao Le, Tai Nguyen, Anh-Tien Nguyen, TrungTin Nguyen, Nhat Ho, Pengtao Xie, Roger Wattenhofer, et al. Exgra-med: Extended context graph alignment for medical vision-language models. *Advances in Neural Information Processing Systems*, 38:109568–109611, 2026.

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.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

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.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1713–1721, 2015.

Szymon Płotka, Gizem Mert, Maciej Chrabaszcz, Ewa Szczurek, and Arkadiusz Sitek. Mamba goes home: Hierarchical soft mixture-of-experts for 3d medical image segmentation. *Advances in Neural Information Processing Systems*, 38:97871–97909, 2026.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, et al. Voxtell: Free-text promptable universal 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37538–37557, 2026.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 24094–24107, 2025.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

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, 2018a.

Yun Wang, Juncheng Li, and Florian Metze. Comparing the max and noisy-or pooling functions in multiple instance learning for weakly supervised sequence learning tasks. *arXiv preprint arXiv:1804.01146*, 2018b.

Ziao Wang, Sixing Yan, Kejing Yin, Xiaofeng Zhang, and William K Cheung. Curv: coherent uncertainty-aware reasoning in vision-language models for x-ray report generation. *Advances in Neural Information Processing Systems*, 38:93686–93713, 2026.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

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.

Yuexin Wu, I-Chan Huang, and Xiaolei Huang. Token imbalance adaptation for radiology report generation. In *Conference on Health, Inference, and Learning*, pp. 72–85. PMLR, 2023.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, et al. Radeval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 546–557, 2025.

Sixing Yan, Ziao Wang, Kejing Yin, William Kwok-wai Cheung, Ka Chun Cheung, and Simon See. Learning self-critiquing mechanisms for region-guided chest x-ray report generation. In *International Conference on Learning Representations*, volume 2026, pp. 78368–78393, 2026.

Zefan Yang, Ge Wang, James Hendler, Mannudeep K Kalra, and Pingkun Yan. X-win: Building chest radiograph world model via predictive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6920–6930, 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.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

Chenhui Zhao, Yiwei Lyu, Asadur Zaman Chowdury, Edward S Harake, Akhil Kondepudi, Akshay T Rao, Xinhai Hou, Honglak Lee, and Todd C Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Jiaheng Zhou, Wei Fang, Luyuan Xie, Yanfeng Zhou, Lianyan Xu, Minfeng Xu, Ge Yang, and Yuxing Tang. Dual-res tandem mamba-3d: Bilateral breast lesion detection and classification on non-contrast chest ct. *Advances in Neural Information Processing Systems*, 38:25494–25527, 2026a.

Qin Zhou, Guoyan Liang, Qianyi Yang, Jingyuan Chen, Sai Wu, Chang Yao, and Zhe Wang. Enhancing reinforcement learning for radiology report generation with evidence-aware rewards and self-correcting preference learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37044–37056, 2026b.

Yang Zhou, Tan L Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick S Goh. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. *Advances in Neural Information Processing Systems*, 37:6625–6647, 2024.

Appendix

Table of Contents

A Data **16**
A.1 Training set 16
A.2 Test set 16

B Report preprocessing **16**
B.1 Report preprocessing example 18

C Diagnostic supervision construction **19**

D Diagnostic loss illustrations **20**

E Experimental setup **21**
E.1 Implementation details 21
E.2 Evaluation metrics 21

F Robustness across training seeds **25**
F.1 Variation across individual runs 26

G Ablation study details **26**

A DATA

We evaluate Diagnostic Loss on abdominal CT data. Table 4 summarizes the datasets used for training and testing. Section A.1 describes the training data, and Section A.2 describes the test data.

Table 4: **Overview of the 3D datasets used for training or testing.** We provide the number of scans, exams, patients, and training and test splits for each dataset.

Dataset	Scans	Exams	Patients	Train	Test
Merlin	25,494	25,494	18,317	20,369	5,125
CancerVerse	24,168	22,867	13,863	24,168	-
In-house	2,706	2,706	1,948	-	2,706
Total	**52,368**	**50,767**	**34,128**	**44,537**	**7,831**

A.1 TRAINING SET

For the 3D experiments reported here, we train models with and without Diagnostic Loss using Merlin (Blankemeier et al., 2026) and CancerVerse (Li et al., 2026b).

A.1.0.1 Merlin. The training set consists of 20,369 contrast-enhanced CT scans. Following Blankemeier et al. (2026), we split the dataset for training (15,309 CTs), and validation (5,060 CTs) sets. Each CT exam is paired with the corresponding radiology report. The dataset includes abdominal and pelvis CT exams conducted between 2012 and 2018 at the Stanford Hospital Emergency Departments.

A.1.0.2 CancerVerse. The training set consists of 24,168 multi-phase CT scans from 13,863 patients[1]. We use all cases as a training set. Each CT exam is paired with the corresponding radiology report. The dataset includes abdominal and pelvis CT scans acquired at the University Hospital Basel, Switzerland and Medipol University Hospital, Türkiye.

A.2 TEST SET

A.2.0.1 Merlin. We use an official test set (Blankemeier et al., 2026), which consists of 5,125 contrast-enhanced CT scans.

A.2.0.2 In-house. The dataset consists of 2,706 multi-center, contrast-enhanced CT scans from 1,948 patients. Each CT examination is paired with the corresponding radiology report. All scans were acquired across 300 sites within a nationwide cohort. We have obtained IRB for this data. *Please note, due to double-blind submission policy, we will add IRB number, if accepted.*

B REPORT PREPROCESSING

Figure 3 shows a free-text CT report and an abbreviated structured output from Qwen3.

[1]CancerVerse dataset is publicly available at https://huggingface.co/datasets/BodyMaps/CancerVerse

Free-text CT report

FINDINGS: Lower thorax: Increased attenuation within the lung bases with vascular crowding is likely secondary to atelectasis. Liver and biliary tree: Cirrhotic liver morphology with nodularity of the liver surface and relative enlargement of the lateral segment. An exophytic enhancing lesion projecting from segment 7 measures 3.5 x 3.1 cm. Another exophytic enhancing lesion along the falciform ligament measures 2.1 x 2.0 cm. In segment 2, there is a 2.0 x 1.3 cm enhancing lesion. Gallbladder: Intraluminal gallstones with moderate wall thickening. Spleen: Status post splenectomy with multiple splenules and peripheral calcification. Additional enhancing nodules in the right pelvis up to 1.9 cm and small enhancing nodules in the omentum up to 1.2 cm. Pancreas: 0.5 cm cystic lesion in the tail; main duct not dilated. Adrenals: Normal. Kidneys: No hydronephrosis; left kidney with subcentimeter hypodensities. Small bowel: Mildly dilated fluid-filled loops up to 3.1 cm with collapsed distal loops. Colon: Sigmoid diverticulosis; submucosal edema in ascending and transverse colon. Bladder: Moderate wall thickening. Prostate: Normal.



Structured output


```
{
  "study_id": "AC4243f52",
  "cleaned_report": "[... The full cleaned report text is here ...]",
  "organs": {
    "Spleen": "Status post splenectomy with multiple splenules. Peripheral
    calcification around several splenules. Enhancing nodules in the right pelvis
    up to 1.9 cm. Small enhancing omental nodules up to 1.2 cm.",
    "Kidney right": "Unremarkable.",
    "Kidney left": "Subcentimeter hypodensities in the left kidney.",
    "Gallbladder": "Intraluminal gallstones with moderate gallbladder wall
    thickening.",
    "Liver": "Cirrhotic morphology with liver surface nodularity and relative
    enlargement of the lateral segment. Exophytic enhancing lesion in segment 7
    measuring 3.5 × 3.1 cm. Additional enhancing lesion along the falciform
    ligament measuring 2.1 × 2.0 cm. Enhancing lesion in segment 2 measuring
    2.0 × 1.3 cm.",
    "Stomach": "Unremarkable.",
    "Pancreas": "0.5 cm cystic lesion in the pancreatic tail. Main pancreatic duct
    not dilated.",
    "Adrenal gland right": "Normal.",
    "Adrenal gland left": "Normal.",
    "Small bowel": "Mildly dilated and fluid-filled loops up to 3.1 cm with
    collapsed distal loops. Wall enhancement not well appreciated.",
    "Duodenum": "Unremarkable.",
    "Colon": "Diverticulosis of the sigmoid colon. Prominent submucosal edema in
    the ascending and transverse colon.",
    "Urinary bladder": "Moderate bladder wall thickening.",
    "Prostate": "Normal."
  }
}
```


Figure 3: **Example of report-to-structured-text conversion.** Qwen3 maps a free-text CT report to a cleaned report and organ-specific findings. The structured output is abbreviated for display.

B.1 REPORT PREPROCESSING EXAMPLE

Figure 4 shows how one report is cleaned and separated into organ-specific texts and labels.

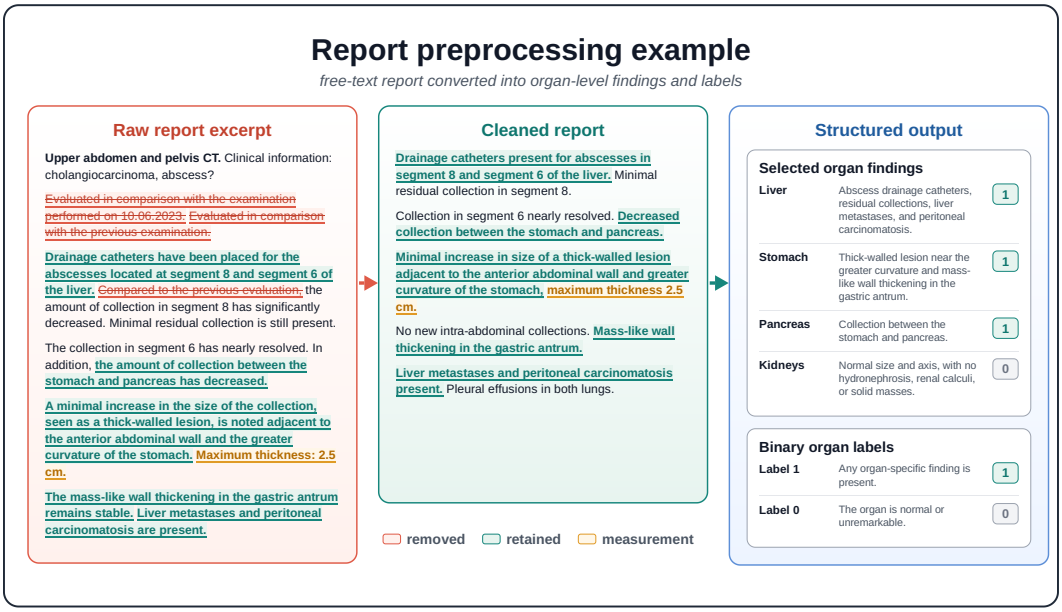


Figure 4: **Example of radiology report preprocessing for a single CT study.** The red panel shows the raw report excerpt, the green panel shows the cleaned current-study report text, and the blue panel shows selected organ-specific outputs and binary organ labels. Text highlights indicate removed phrasing, retained findings, and retained measurements. Only selected organs are shown for readability; the full preprocessing output contains all 11 organ groups.

C DIAGNOSTIC SUPERVISION CONSTRUCTION

Figure 5 illustrates how supported and unsupported diagnostic targets are constructed from an organ-specific reference text.

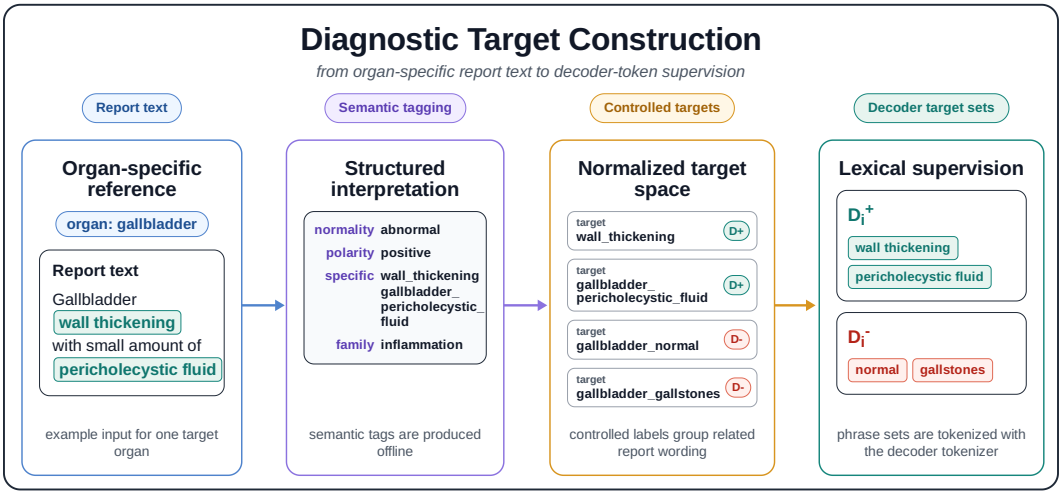


Figure 5: **Schematic overview of diagnostic target construction for the diagnostic loss.** An organ-specific reference text is assigned structured semantic tags, mapped into a controlled diagnostic target space, and converted into supported and unsupported lexical target sets for the decoder loss. Supported targets are findings stated in the reference text; unsupported targets are alternatives for the same organ that are discouraged during training. The gallbladder example displays representative targets and lexical forms for one organ.

D DIAGNOSTIC LOSS ILLUSTRATIONS

Figure 6 contrasts the two training objectives.

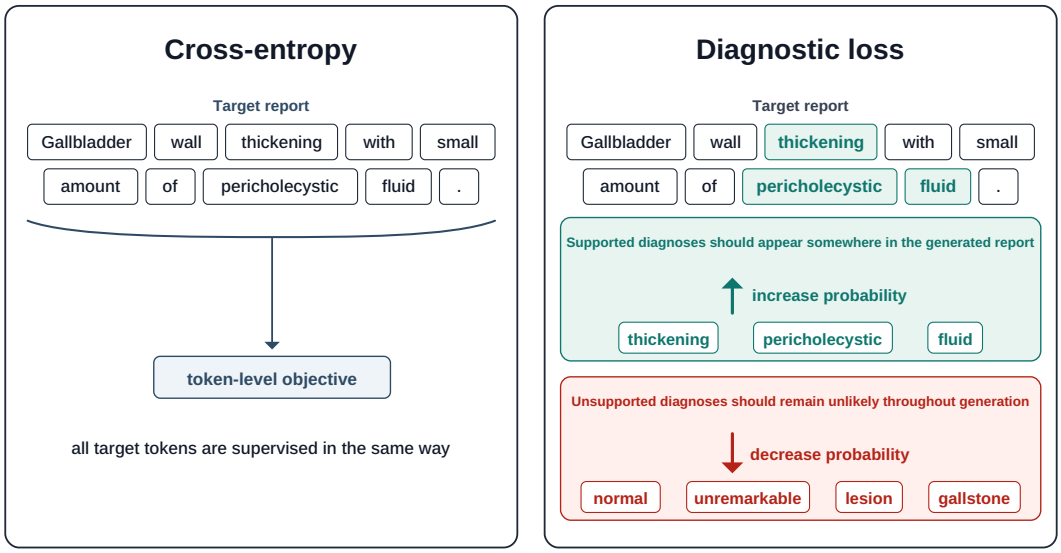


Figure 6: **Conceptual comparison between token-level cross-entropy and the proposed *diagnostic loss*.** Cross-entropy supervises each next-token prediction in the reference radiology report. Diagnostic loss increases probability assigned to supported targets at one or more decoder positions and suppresses probability assigned to unsupported targets across positions. For readability, tokens are shown as whole words, although model tokenizers may split words into smaller units.

E EXPERIMENTAL SETUP

E.1 IMPLEMENTATION DETAILS

We implement our method using PyTorch and train on up to four NVIDIA GH200 GPUs. During report-generation training, visual features are fixed. We train low-rank adaptation (LoRA) parameters in each language decoder together with its visual-to-text projector or adapter, using the same visual initialization for the cross-entropy and Diagnostic Loss runs of each backbone. Merlin and RadFM use released visual weights, SPECTRE uses a pretrained visual checkpoint, and fVLM, BTB3D, and HLIP use locally trained visual checkpoints.

Merlin uses the RadLLaMA-7B decoder; BTB3D uses Llama-3.1-8B-Instruct; fVLM, HLIP, and SPECTRE use Qwen2.5-0.5B; and RadFM uses MedLLaMA-13B. Each report generator predicts one organ finding at a time from an organ prompt. For each backbone, the cross-entropy and Diagnostic Loss runs use the same decoder, tokenizer, visual initialization, and pooled Merlin–CancerVerse report-generation training data.

The full decoder weights remain frozen. Merlin uses LoRA rank 512, alpha 16, and dropout 0.1 on the decoder attention query and value projections. BTB3D uses rank 128, alpha 256, and dropout 0.05; fVLM, HLIP, RadFM, and SPECTRE use rank 16, alpha 32, and dropout 0.05. These latter five models apply LoRA to the decoder query, key, value, and output projections and to the gate, up, and down projections of its feed-forward layers. For each backbone, the cross-entropy and Diagnostic Loss runs use the same LoRA configuration and trainable projector or adapter.

Diagnostic Loss settings. The backbone-specific loss coefficients in the main experiments were selected through a small grid search and are listed in Table 5. All runs use $\tau = 8$.

Table 5: **Diagnostic Loss coefficients in the main experiments.** The columns correspond to the supported- and unsupported-target coefficients in Eq. 9.

Backbone	λ_+	λ_-
Merlin–RadLLaMA	0.001	0.0005
fVLM–Qwen	0.001	0.00025
BTB3D	0.001	0.00025
HLIP–Qwen	0.001	0.00025
RadFM-v3	0.01	0.01
SPECTRE–Qwen	0.01	0.005

E.2 EVALUATION METRICS

E.2.1 LEXICAL METRICS

The lexical metrics used in the experiments compare generated and reference texts through shared tokens, ordered word sequences, or weighted phrase overlap. Most of them operate on *n*-grams, meaning contiguous groups of n tokens: unigrams contain one token, bigrams contain two adjacent tokens, and longer *n*-grams capture longer local phrases. They are computed using the Microsoft Common Objects in Context (MS COCO) caption-evaluation setup (Chen et al., 2015). The equations below state the standard nondegenerate definitions. In the implementation, zero modified precisions and empty generated texts are handled by the small constants and brevity penalty used in the COCO caption-evaluation scorer, rather than by passing zero directly to the logarithm. For METEOR, examples with no matched unigrams receive zero score, so the chunk penalty is only relevant when at least one unigram is matched. For CIDEr, cosine-similarity terms with a zero-norm TF-IDF vector are treated as zero.

BLEU-4 (Papineni et al., 2002) measures how many generated *n*-grams also appear in the reference. Plain precision would allow a model to repeat a correct word many times and receive excessive credit, so BLEU uses clipped counts: each generated *n*-gram can be counted only up to its frequency in the reference. For a generated text c_i and a reference text r_i, the modified precision for order n is:

$$p_n = \frac{\sum_{g \in G_n(c_i)} \min\left(\text{count}_{c_i}(g), \text{count}_{r_i}(g)\right)}{\sum_{g \in G_n(c_i)} \text{count}_{c_i}(g)}. \tag{11}$$

BLEU-4 combines the modified precisions for $n = 1, \ldots, 4$ using a geometric mean and applies a brevity penalty:

$$\mathrm{BLEU}_4 = \mathrm{BP} \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right), \tag{12}$$

$$\mathrm{BP} = \begin{cases} 1, & |c_i| > |r_i|, \\ \exp\left(1 - \frac{|r_i|}{|c_i|}\right), & |c_i| \leq |r_i|. \end{cases} \tag{13}$$

Here $|c_i|$ and $|r_i|$ denote text lengths after metric tokenization. The brevity penalty reduces the score when the generated text is shorter than the reference, because high precision alone can otherwise be obtained by generating only a small subset of the radiology report. BLEU is therefore sensitive to how much of the generated wording is supported by the reference, but it is less direct about missing reference content.

ROUGE-L (Lin, 2004) measures overlap through the longest common subsequence (LCS) between the generated and reference texts. A subsequence preserves word order but does not require the matched words to be adjacent. If L_i is the LCS length, ROUGE-L defines an LCS-based precision and recall:

$$P_L = \frac{L_i}{|c_i|}, \qquad R_L = \frac{L_i}{|r_i|}. \tag{14}$$

These quantities are combined with an F-measure:

$$\text{ROUGE-L} = \frac{(1+\beta^2) P_L R_L}{R_L + \beta^2 P_L}. \tag{15}$$

The parameter β controls the precision–recall trade-off. This makes ROUGE-L more sensitive to reference coverage than BLEU, because it rewards generated texts that preserve longer ordered sequences from the reference.

METEOR (Lavie & Denkowski, 2009) starts from unigram alignment rather than higher-order *n*-gram matching. The generated and reference texts are aligned using exact matches, stems, and synonyms, so two wording variants can receive credit even when their surface forms are not identical. If m_i tokens are matched, the METEOR precision, recall, and weighted harmonic mean are:

$$P_M = \frac{m_i}{|c_i|}, \qquad R_M = \frac{m_i}{|r_i|}, \qquad F_{\mathrm{mean}} = \frac{10 P_M R_M}{R_M + 9 P_M}. \tag{16}$$

This weighted harmonic mean gives more weight to recall than to precision. METEOR then checks whether the matched words form long contiguous runs or are split across the sentence. Let q_i be the number of matched chunks and let m_i be the number of matched unigrams. The penalty has the form:

$$\text{Penalty} = \gamma \left(\frac{q_i}{m_i}\right)^{\theta}, \tag{17}$$

where γ and θ are fixed metric parameters from the METEOR implementation. The final score is:

$$\text{METEOR} = (1 - \text{Penalty}) F_{\mathrm{mean}}. \tag{18}$$

This gives the metric a different behavior from pure overlap counts, because some wording variants can receive credit when they are aligned by the matching modules.

CIDEr (Vedantam et al., 2015) also compares *n*-gram overlap, but it weights *n*-grams by how informative they are in the evaluation collection. Common phrases receive less weight, while rarer phrases receive more weight. For an *n*-gram g, the term frequency in text x is:

$$\mathrm{tf}_{n,g}(x) = \frac{\mathrm{count}_x(g)}{\sum_{g' \in G_n(x)} \mathrm{count}_x(g')}. \tag{19}$$

Let $\mathrm{df}_{n,g}$ be the number of reference texts in the evaluation set that contain g. The inverse document frequency is:

$$\mathrm{idf}_{n,g} = \log\left(\frac{N}{\mathrm{df}_{n,g}}\right), \tag{20}$$

with standard smoothing in the implementation to avoid undefined values. The term-frequency–inverse-document-frequency (TF-IDF) vector for order n has components:

$$v_{n,g}(x) = \mathrm{tf}_{n,g}(x)\mathrm{idf}_{n,g}. \tag{21}$$

Let $\boldsymbol{v}_n(c_i)$ and $\boldsymbol{v}_n(r_i)$ be the resulting vectors for the generated and reference texts. The score for one n-gram order is their cosine similarity:

$$\mathrm{CIDEr}_n(c_i, r_i) = \frac{\boldsymbol{v}_n(c_i) \cdot \boldsymbol{v}_n(r_i)}{\|\boldsymbol{v}_n(c_i)\| \, \|\boldsymbol{v}_n(r_i)\|}. \tag{22}$$

The final CIDEr score averages the first four n-gram orders:

$$\mathrm{CIDEr}(c_i, r_i) = \frac{10}{4} \sum_{n=1}^{4} \mathrm{CIDEr}_n(c_i, r_i). \tag{23}$$

The factor 10 is the conventional CIDEr scaling used by the metric definition. Compared with unweighted overlap, CIDEr gives more influence to phrases that are informative in the evaluation collection.

E.2.2 Clinical-content metrics

Clinical-content metrics make the distinction between lexical and clinical evaluation operational by comparing radiology report meaning instead of only surface wording. This matters for diagnostic loss because the loss is designed to affect diagnostic expressions derived from the reference report in the decoder distribution. The experiments compute RadGraph and GREEN, or Generative Radiology Report Evaluation and Error Notation (Ostmeier et al., 2024), with the RadEval framework (Xu et al., 2025). These metrics compare generated and reference reports automatically; they are not a radiologist review tailored to individual CT organs and do not establish clinical safety on their own.

RadGraph (Jain et al., 2021) evaluates radiology reports by applying an automatic clinical information-extraction model to both the generated and reference texts. The RadGraph paper defines an annotation schema for radiology reports and trains a benchmark extractor to predict this schema from free text. The extractor first identifies text spans that correspond to anatomy and observations, then assigns observation status, such as definitely present, uncertain, or definitely absent, and predicts relations between entity pairs. The benchmark model reported by RadGraph uses a transformer-based entity and relation extraction architecture with biomedical language-model initialization. A generated radiology report therefore receives credit only when the extracted clinical structure agrees with the reference: what observation is mentioned, where it is located, and whether it is asserted or negated. RadEval returns simple, partial, and complete RadGraph variants; the experiments report the complete entity-and-relation score as $\mathrm{F1}_{\mathrm{RG}}$. If $T(c_i)$ and $T(r_i)$ denote the entity-and-relation tuples extracted from the generated and reference texts, RadGraph precision and recall are:

$$P_{\mathrm{RG}} = \frac{|T(c_i) \cap T(r_i)|}{|T(c_i)|}, \qquad R_{\mathrm{RG}} = \frac{|T(c_i) \cap T(r_i)|}{|T(r_i)|}, \tag{24}$$

where the intersection counts extracted tuples that match between the generated and reference radiology reports. The complete RadGraph F1 score is:

$$\mathrm{F1}_{\mathrm{RG}} = \frac{2P_{\mathrm{RG}}R_{\mathrm{RG}}}{P_{\mathrm{RG}} + R_{\mathrm{RG}}}. \tag{25}$$

If either report is empty or either extracted graph has no entities, the corresponding RadGraph score contribution is treated as zero by the RadEval implementation. The reported value is the mean $\mathrm{F1}_{\mathrm{RG}}$ over all evaluated organ-specific examples.

GREEN uses a generative language model as the evaluator. The evaluator receives a reference radiology report and a candidate radiology report, then produces a structured error analysis rather than only a single similarity value. The original GREEN method uses GPT-4 analyzes of synthetic radiology report pairs to train radiology-domain language models that identify and classify report errors. These errors include false findings, missing findings, wrong anatomical location, wrong severity, and unsupported comparison statements. GREEN also distinguishes clinically significant

errors from less important discrepancies, which makes the score closer to an error-based clinical review than to lexical overlap. In the experiments, RadEval computes GREEN with the released local GREEN-RadLLaMA2-7b evaluator rather than with an API-based GPT-4 judge.

The scalar GREEN score is the aggregation of the structured evaluator output used in the evaluation implementation. Let M_i be the number of matched findings in example i, and let $E_{i,k}^{\mathrm{sig}}$ be the number of clinically significant errors in error category k. The per-example score is:

$$s_i^{\mathrm{GREEN}} = \begin{cases} \dfrac{M_i}{M_i + \sum_k E_{i,k}^{\mathrm{sig}}}, & M_i > 0, \\ 0, & M_i = 0. \end{cases} \tag{26}$$

This expression summarizes the structured GREEN error output as a scalar score: matched findings increase the numerator, while clinically significant errors increase the denominator. Higher values, therefore, indicate that the generated radiology report preserves more reference findings while introducing fewer clinically significant errors. The reported GREEN value is the mean over the evaluated examples:

$$\mathrm{GREEN} = \frac{1}{N} \sum_{i=1}^{N} s_i^{\mathrm{GREEN}}. \tag{27}$$

F ROBUSTNESS ACROSS TRAINING SEEDS

Table 1 shows gains from Diagnostic Loss across several backbones, with one training run per setting. Repeating every backbone across many seeds would require substantial additional training. We therefore complement the cross-backbone comparison with nine matched BTB3D runs on Merlin. This experiment tests whether the BTB3D gains persist across training seeds and resampled test studies.

We compared BTB3D trained with cross-entropy (CE) and with CE plus Diagnostic Loss (DLoss) using nine matched random seeds: 7, 13, 42, 101, 103, 107, 2026, 2718, and 31415. Each model was evaluated on the full Merlin test set of 5,125 studies and 56,375 organ-specific report rows. CE and DLoss runs with the same seed use the same test rows. All rows, including empty generations, remain in the metric denominators.

Statistical analysis. For each metric, we subtract the CE score from the DLoss score within each matched seed pair, then average the nine differences. A positive mean effect favors DLoss. Table 6 reports this effect and two bootstrap 95% confidence intervals, each based on 5,000 resamples. The seed bootstrap samples the nine CE–DLoss seed pairs with replacement and recomputes the mean effect. The seed–study bootstrap also samples test studies with replacement. It retains all 11 organ rows from each sampled study and uses the same sampled studies for CE and DLoss in every selected seed pair. Each interval spans the 2.5th to 97.5th percentiles of its resampled effects. Pairing preserves the comparison within a training seed; sampling by study respects the clustering of organ rows. For corpus BLEU, we recompute the score from n-gram counts in each resample. Other metrics use saved row scores, with the original test-corpus inverse document frequencies held fixed for CIDEr. The tables report effects on the native scales of the saved metric outputs.

Table 6: **Paired BTB3D results on the full Merlin test set.** Each effect is the mean DLoss-minus-CE difference over nine matched seeds. Intervals are 95% confidence intervals; positive values favor DLoss. Point estimates use all 56,375 organ rows per run.

Metric	Mean effect	Seed bootstrap	Seed–study bootstrap
BLEU-1	$+0.0616$	$[+0.0046, +0.1242]$	$[+0.0050, +0.1238]$
BLEU-2	$+0.0698$	$[+0.0099, +0.1387]$	$[+0.0096, +0.1385]$
BLEU-3	$+0.0738$	$[+0.0114, +0.1460]$	$[+0.0116, +0.1450]$
BLEU-4	$+0.0752$	$[+0.0124, +0.1483]$	$[+0.0127, +0.1475]$
METEOR	$+0.0795$	$[+0.0169, +0.1480]$	$[+0.0164, +0.1483]$
ROUGE-L	$+0.0689$	$[+0.0138, +0.1275]$	$[+0.0137, +0.1277]$
CIDEr	$+0.7834$	$[+0.1528, +1.4527]$	$[+0.1539, +1.4560]$
GREEN	$+0.0455$	$[+0.0197, +0.0728]$	$[+0.0197, +0.0732]$
RadGraph Complete	$+0.0795$	$[+0.0203, +0.1447]$	$[+0.0203, +0.1451]$

GREEN and RadGraph Complete have positive mean differences, and both reported bootstrap intervals lie above zero. The seven lexical metrics have positive mean differences and positive bootstrap intervals as well. Table 7 shows variation across runs: GREEN and RadGraph Complete improve in eight of nine seed pairs, while BLEU-4 improves in seven of nine. The positive mean effects do not imply improvement for every trained model.

F.1 Variation across individual runs

Table 7 shows the within-seed differences for BLEU-4 and the two clinical-content metrics. It makes the spread across runs visible alongside the average effects and intervals above.

Table 7: **Within-seed effects for representative metrics.** Each entry is the DLoss-minus-CE difference on the full Merlin test set, in the native metric scale. These are paired training runs, not independent test-row samples.

Seed	BLEU-4	GREEN	RadGraph Complete
7	+0.0617	+0.0726	+0.0371
13	+0.2790	+0.0913	+0.2497
42	+0.0267	+0.0433	+0.0689
101	+0.0861	+0.0205	+0.0918
103	+0.0045	+0.0141	+0.0165
107	+0.2391	+0.1216	+0.2406
2026	−0.0449	+0.0307	−0.0611
2718	−0.0225	−0.0215	+0.0184
31415	+0.0472	+0.0372	+0.0538

The bootstrap intervals are unadjusted for multiple metrics. With nine seed pairs, they describe uncertainty for this BTB3D protocol on Merlin; they do not establish seed robustness for the other backbones or the external test cohort.

G Ablation study details

Table 3 compares matched fVLM runs with seed 13 on the full Merlin test set of 5,125 studies and 56,375 organ findings. All percentage changes use the cross-entropy (CE) run from this ablation as their baseline. Checkpoints are selected by validation CE. S-only adds the supported-target loss to CE; U-only adds the unsupported-target loss to CE. The pooling variants add both diagnostic components to CE and change one pooling operator. The ablations used two NVIDIA A100 GPUs (40 GB each), whereas the main experiments used NVIDIA GH200 GPUs.

The S-max and S-mean variants replace supported-target noisy-OR in Eq. 5 with

$$q_{i,d}^{\text{S-max}} = \max_{t \in \mathcal{P}_i} m_{i,d,t}, \qquad q_{i,d}^{\text{S-mean}} = \frac{1}{|\mathcal{P}_i|} \sum_{t \in \mathcal{P}_i} m_{i,d,t}. \tag{28}$$

Both use $-\log q_{i,d}$ as in Eq. 6. U-mean instead replaces Eq. 7 with

$$L_{i,d}^{-,\text{U-mean}} = \log \left(\frac{1}{|\mathcal{P}_i|} \sum_{t \in \mathcal{P}_i} m_{i,d,t} \right). \tag{29}$$

Thus, the logarithm is applied after averaging the unsupported-target probability masses. This variant does not use τ; supported-target pooling remains noisy-OR. As in the main equations, probability clamping is omitted for readability.

Full DLoss improves all three reported metrics over the matched CE run, whereas either component alone lowers them. For supported targets, maximum pooling discards evidence outside the strongest position, while a mean can dilute evidence localized to a few positions; both alternatives score below noisy-OR here. U-mean has the largest declines, consistent with the need to penalize high unsupported-target probability even when it occurs at a single position. This run also generates 1,423 empty findings (2.52%), compared with 48 (0.09%) for CE; all are included in the metric denominators. Together, these fVLM ablations support the two-component objective and its asymmetric pooling rules.