CURE: Curriculum-guided Multi-task Training for Reliable Anatomy Grounded Report Generation

Supplementary Material

6. Implementation and Training Details

This section provides a detailed overview of the experimental setup, including hardware, hyperparameters, and the curriculum learning configuration to ensure full reproducibility.

6.1. Model and Hardware Setup

Our framework is implemented in PyTorch [34] using the Hugging Face ecosystem [52], particularly the `SFTTrainer` from the TRL library for supervised fine-tuning. All experiments were conducted within a SLURM-managed High-Performance Computing (HPC) cluster equipped with NVIDIA RTX A6000 GPUs, each providing 48 GB of VRAM.

Each training run for our final model was executed on a single GPU with 60 GB of system memory RAM. The total training time for the final 9000-step CURE model was approximately 45 hours, comprising a 15-hour pre-training stage (3000 steps) followed by a 30-hour multi-task fine-tuning stage (6000 steps).

6.2. Hyperparameter Details

The complete training process for the final version of CURE spans 9000 steps and is strictly divided into two main phases. For clarity, the chronological pipeline is structured as follows:

- **Phase 1: Pre-training (3000 steps).** The model is trained exclusively on the Chest ImaGenome dataset.
- **Phase 2: Multi-task Fine-tuning (6000 steps).** The optimizer and scheduler states are reset, preserving only the model weights from Phase 1. This phase applies our curriculum learning framework and is further split into two stages:
  - *Stage 2a: Warm-up (3000 steps).* Uniform sampling is applied across all datasets and tasks.
  - *Stage 2b: Cyclic Re-weighting (3000 steps).* A performance evaluation dictates new sampling weights, which are then fixed for these concluding steps.

Key hyperparameters, which remained consistent across all stages unless otherwise noted, are detailed in Table 8.

6.3. Curriculum Learning Details

Our curriculum learning framework is applied during the multi-task fine-tuning phase (Phase 2) to dynamically prioritize underperforming tasks. The protocol operates via two conceptual mechanisms:

Table 8. **Hyperparameter Configuration.** Detailed hyperparameters for the pre-training and multi-task fine-tuning stages.

Hyperparameter	Value
Model & Training	
Base Model	MedGemma-4B-IT
Quantization	4-bit NF4
Precision	BF16
Optimizer	Fused AdamW
Learning Rate	2×10^{-4}
LR Scheduler	Linear
Warmup Ratio	0.03
Batch Size (per device)	5
Gradient Accum. Steps	5
Effective Batch Size	25
Max Grad Norm	0.3
LoRA Configuration	
Rank (r)	16
Alpha (α)	32
Dropout	0.05
Target Modules	All linear layers
Modules to Save	`lm_head, embed_tokens`

1. **Warm-up (Stage 2a):** By sampling all datasets and intra-dataset categories uniformly, the model receives balanced exposure to all tasks. This establishes a stable performance baseline before adaptation begins.
2. **Error-Aware Re-weighting (Stage 2b):** The model's performance is evaluated on validation sets to recalculate sampling weights, forcing the network to focus on its most frequent errors.

While our framework supports continuous cyclic re-weighting (e.g., recalculating weights every M steps), our ablation studies (Table 7) showed that a single weight update—calculated immediately after the warm-up and fixed for the remainder of training—yielded the most effective and stable configuration for our final **CURE** model.

The re-weighting mechanism itself operates at two levels of granularity:

Curriculum Scoring & Re-weighting. Our curriculum operates at two levels: **inter-dataset** (across different data sources) and **intra-dataset** (across fine-grained categories, such as the 8 phrase classes in MS-CXR). For both levels, we compute an aggregate performance score s using a task-

adaptive metric:

$$s = \alpha \cdot \text{IoU} + (1 - \alpha) \cdot \text{CXRFEScore}. \quad (3)$$

The weight α adapts to the subtask requirements: $\alpha = 0$ for text-only generation (e.g., AGRG "Describe"), $\alpha = 1$ for pure localization (e.g., AGRG "Locate"), and $\alpha = 0.8$ when both modalities are evaluated. The resulting error $e = 1 - s$ updates the sampling probabilities at both the dataset and category levels, directing the model toward its most challenging concepts.

To provide a more granular visualization of the curriculum's adaptive mechanism, Figures 4 and 5 illustrate the weight evolution from an experiment with a more frequent re-weighting schedule (every 500 steps). While this specific timing differs from our final CURE model, these plots clearly demonstrate the dynamic nature of the framework in action.

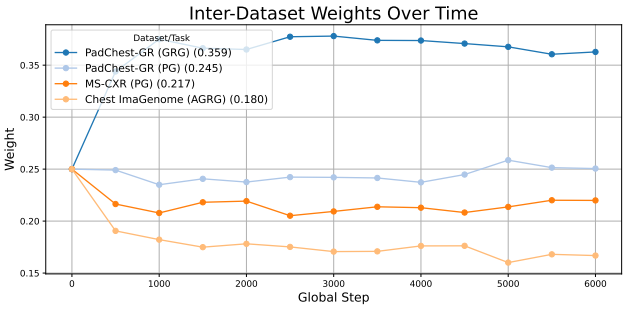


Figure 4. **Visualization of Inter-Dataset Weight Dynamics.** This plot illustrates the curriculum's adaptation from an experiment with frequent updates (every 500 steps). It shows how sampling probabilities for each data source evolve over time in response to the model's performance.

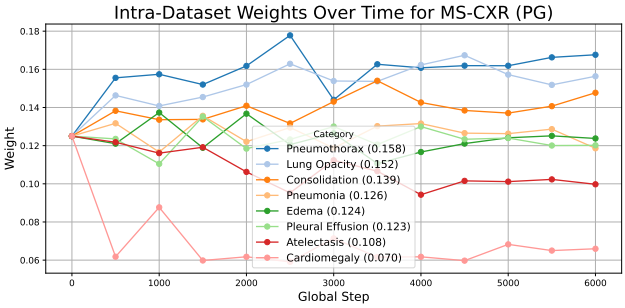


Figure 5. **Visualization of Intra-Dataset Weight Dynamics for MS-CXR.** This plot shows the category-level weight evolution for the 8 phrase classes in MS-CXR, taken from the same experiment with updates every 500 steps. The weights are periodically adjusted to prioritize classes with higher error rates.

7. Datasets and Task Formulation

7.1. Detailed Task I/O Formats

Table 9 provides representative examples of the instructional prompts and expected output formats for each dataset and task used during training and evaluation.

7.2. Dataset Preprocessing and Splits

We used the official train, validation, and test splits provided by each dataset. All images were processed using a custom pipeline built with the Albumentations library [7] and resized to a final resolution of 448×448 pixels. The specific image transformations varied between the training and validation/test phases to ensure data diversity during training and deterministic evaluation.

Training Pipeline. The training pipeline is stochastic, designed to improve model robustness to variations in X-ray acquisition. For each training image, the following transformations are applied:

- **CLAHE:** Applied with a 50% probability to simulate varying contrast levels, using a random clip limit uniformly sampled between 1.0 and 4.0, and a fixed tile grid size of $(8, 8)$.
- **Spatial Augmentations:** Spatial transformations included random resized cropping (30% probability) and affine transformations (translation, scaling up to $\pm 10\%$, and rotation up to $\pm 15°$) applied with a 50% probability. Horizontal flipping was disabled due to the inherent left-right asymmetry of thoracic anatomy. Color-based augmentations such as jitter or Gaussian noise were explicitly disabled.
- **Regularization:** To improve stability, 30% of the training samples bypassed the spatial augmentations and instead used the deterministic validation pipeline described below.

Crucially, the entire training pipeline is bounding box-aware. When spatial transformations are applied, Albumentations simultaneously transforms the corresponding bounding box coordinates. To preserve alignment between the visual and textual modalities, the ground-truth text supervision provided to the MedGemma model is dynamically updated to reflect the augmented coordinates before training. This ensures that every augmented image remains correctly paired with its corresponding, spatially consistent text supervision.

Validation and Test Pipeline. The validation and test pipelines are deterministic. Unlike the stochastic training pipeline, these splits utilized Contrast-Limited Adaptive Histogram Equalization (CLAHE) as a fixed preprocessing normalization step rather than an augmentation. Given the high dynamic range and variable exposure settings inherent to Chest X-rays, applying deterministic CLAHE (clip limit 3.0, tile grid size $(8, 8)$) standardizes the local contrast dis-

tribution across all evaluation samples. This ensures that fine-grained clinical features—which are often obscured in low-contrast regions—are enhanced consistently for the visual encoder during inference. Finally, images were resized to 448×448 pixels.

7.3. Chest ImaGenome Evaluation Benchmark

We strictly adhered to the official MIMIC-CXR data splits. As detailed in the main text, the Chest ImaGenome dataset provides scene graphs for frontal-view images in MIMIC-CXR, linking anatomical bounding boxes to textual descriptions. While we utilized the original Chest ImaGenome annotations (text snippets and bounding boxes) directly for the large-scale training and validation of the Anatomy-Grounded Report Generation (AGRG) task, we devised a rigorous protocol for the test phase to ensure both computational feasibility and high-quality metric calculation.

Computational Constraints and Subsampling. Extracting all valid (image, location) pairs from the scene graphs of the official MIMIC-CXR test set yields a large pool of approximately 123,000 evaluation instances derived from 3,403 unique frontal-view images. While a comprehensive evaluation on the full set is theoretically ideal, it presents significant pragmatic challenges due to the inference latency of large multimodal models.

For example, generating anatomy-grounded reports for 1000 image-location pairs using a fine-tuned MedGemma-4B-IT takes approximately 1 hour and 20 minutes on a single NVIDIA RTX A6000 GPU. Extrapolating this to the full test set results in roughly 164 hours (nearly a week) of continuous inference time for a single model evaluation. To facilitate faster experimentation without sacrificing statistical rigor, we curated a representative, stratified subset of 1000 samples. This subset served as a fixed artifact, ensuring that all models in our experiments were evaluated on the same diverse set of examples.

Generating High-Quality Textual Ground Truth for Evaluation. For the test subset, we sought to improve the granularity and quality of the textual ground truth. The original Chest ImaGenome dataset employs a pre-LLM NLP pipeline to associate radiology report snippets with anatomical locations. While sufficient for large-scale training, these snippets can be somewhat noisy, as they consist of raw fragments from the original radiology report and may include details not strictly related to the specified anatomical location. Moreover, they were not generated using modern language models capable of producing concise, location-tailored phrasing.

To establish a robust reference standard for evaluation metrics, we maintained the original ground-truth bounding boxes but enhanced the textual annotations. We utilized `gemini-2.5-flash-lite` to synthesize concise, location-specific "mini-reports" derived directly from the complete original radiology reports. This ensures the model is evaluated against coherent, radiologist-style descriptions that are explicitly relevant to each anatomical region of interest. The Gemini-generated mini-reports were used only as evaluation ground truth, never for training. The prompt used for this refinement is detailed below:

```
You will be provided with a chest x-ray report and a
     specified anatomical location. Your task is to
     generate a JSON object in the following format:
     {"reasoning": "", "mini-report": ""}

Guidelines:

- reasoning: Begin your reasoning by identifying and
     naming anatomical regions in close proximity to
     the specified location. Then, briefly summarize
     the report as a sequence of findings/observations.
     Lastly, identify all findings relevant to the
     specified location. A finding or observation is
     relevant if it meets any of the following
     criteria: (1) it explicitly describes the
     specified anatomical location; (2) it explicitly
     describes a region anatomically very close to the
     specified location, where the description is
     highly likely to also apply to the specified
     location; (3) it makes a general description from
     which it logically and with absolute certainty
     follows that the description applies to the
     specified location as a specific instance (e.g.,
     "both lungs are clear" implies "the right lung is
     clear"; "no bone abnormalities" implies "the right
     clavicle presents no abnormalities"); or (4) it
     describes devices, tubes, or other objects
     traversing or situated within the specified
     anatomical location. Present your reasoning as a
     single, continuous paragraph, strictly avoiding
     newlines and special characters.
- mini-report: From the relevant information identified
     in your reasoning, synthesize a concise and
     accurate mini-report, written in a style
     consistent with a radiologist's findings,
     specifically detailing the findings related to the
     specified anatomical location.
- If the report contains no findings or descriptions
     pertinent to the specified anatomical location,
     set the value of "mini-report" to "N/A".
- Make sure to use JSON format as shown above.
```

Stratified Sampling Strategy. The full test pool consists of 35,042 image-location pairs that contain descriptive findings and a larger set of image-location pairs annotated only with bounding boxes (i.e., normal or unmentioned regions). To construct the 1000-sample benchmark, we selected 700 instances with descriptive findings and 300 without.

To ensure the subset was representative of the broader test distribution, we applied a stratified sampling strategy. We generated structured annotations for the candidate mini-reports using `gemini-2.5-flash-lite` to label the presence of abnormalities and medical devices:

```
You will be provided with a chest X-ray report or
     sentence. Your task is to analyze the text and
     determine:
```

Table 9. **Summary of Datasets, Tasks, and I/O Formats.** PG = Phrase Grounding; GRG = Grounded Report Generation; AGRG = Anatomy-Grounded Report Generation; RG = Report Generation. The table illustrates representative prompts and output formats for each dataset. Bounding boxes are denoted as `[center_x, center_y, width, height]`. For MIMIC-CXR (evaluation only), reports are generated using the GRG, AGRG, or hybrid AGRG+GRG approaches, and subsequently post-processed to remove bounding box coordinates prior to text-based evaluation.

Dataset	Task	Prompt	Example Output
MS-CXR	PG	Ground the phrase: {*phrase*}	{*phrase*}: [cx_1,cy_1,w_1,h_1] [cx_2,cy_2,w_2,h_2] ...
PadChest-GR	PG	Ground the phrase: {*phrase*}	{*phrase*}: [cx_1,cy_1,w_1,h_1] [cx_2,cy_2,w_2,h_2] ...
	GRG	Generate a grounded report.	Slight residual atelectasis in the right pulmonary base [0.29,0.66,0.18,0.20]. Minimal blunting of the costophrenic angle [0.81,0.74,0.33,0.39].
Chest ImaGenome	AGRG	Locate and describe the {*location*}.	Location of the abdomen: [0.48,0.78,0.73,0.45]. Description: No free air below the right hemidiaphragm is seen.
		Locate the {*location*}.	Location of the right cardiophrenic angle: [0.33,0.71,0.09,0.12].
		Describe the {*location*}.	Description of the left chest wall: Left chest wall pacer defibrillator is unchanged in position.
VinDr-CXR (Eval., zero-shot)	PG	Ground the phrase: {*phrase*}	Cardiomegaly: [0.57,0.65,0.55,0.37]
	GRG	Generate a grounded report.	Consolidation [0.29,0.54,0.18,0.13]. Infiltration [0.27,0.49,0.27,0.24]. Pleural effusion [0.82,0.72,0.25,0.14].
MIMIC-CXR (Eval.)	RG via GRG	Generate a grounded report.	{*Grounded report*}
	RG via AGRG	Locate and describe the {*location*}. ($\times N$ anatomical locations)	{*Report for location 1*} ... {*Report for location N*}
	RG via AGRG + GRG	Combine AGRG and GRG generations.	{*Report for location 1*} ... {*Report for location N*} {*GRG report*}

```
1. Whether any abnormalities or pathologies are
     mentioned.
2. Whether any medical devices or foreign objects are
     mentioned.

Output format:
```

```
Return a JSON object with the following fields:

{
  "reason": "A brief explanation of your reasoning.",
  "mentions_abnormalities": "yes" | "no",
  "mentions_devices": "yes" | "no"
}
```

Using these labels, the 700-sample partition was balanced across anatomical locations, abnormality status, and the presence of medical devices. The 300-sample partition (without specific findings) was sampled uniformly across anatomical locations to preserve anatomical diversity. This procedure yields a balanced evaluation benchmark derived strictly from the official test split.

7.4. VinDr-CXR for Zero-Shot Generalization

To assess model robustness against domain shifts and unseen data distributions, we employ the VinDr-CXR dataset [32] as a zero-shot benchmark.

Dataset Characteristics. VinDr-CXR consists of 15,000 training and 3000 testing frontal-view Chest X-rays. Each image was annotated by a consensus of three radiologists for the presence of 28 common thoracic diseases and findings. These findings are categorized into 22 localizable classes (annotated with bounding boxes) and 6 global classes (image-level labels only).

Zero-Shot Protocol. We exclude the VinDr-CXR training set entirely. None of the models evaluated in this work (including CURE and all baselines) were trained or fine-tuned on any portion of VinDr-CXR. Consequently, all results reported on this dataset reflect pure zero-shot transfer capabilities.

Task Adaptation. Since VinDr-CXR provides structured classification and detection labels rather than narrative radiology reports, we adapted the annotations to align with our text-based generation tasks:

- **Phrase Grounding (PG):** We mapped the short class labels (e.g., "ILD", "Enlarged PA") to full natural language phrases (e.g., "Interstitial lung disease", "Enlarged pulmonary artery"). We generated evaluation instances for every localizable finding present in the test set, resulting in 2,108 zero-shot phrase grounding queries.
- **Grounded Report Generation (GRG):** To create reference targets for report generation, we synthesized "pseudo-reports" from the structured annotations. For a given image, we aggregated all positive findings; localizable findings were converted into text strings containing the finding name followed by their bounding box coordinates (e.g., "Atelectasis [cx, cy, w, h]"), while global findings were appended as text-only sentences. These phrases were concatenated to form a complete, grounded target sequence, allowing us to compute both textual overlap and localization metrics.

8. Evaluation Protocol

8.1. Metric Calculation

All metrics were computed using publicly available official implementations to ensure reproducibility.

CheXbert Metrics. We used the official CheXbert implementation [44], available at https://pypi.org/project/f1chexbert/, to compute clinical correctness metrics. Specifically, we report precision, recall, and F1 scores for all 14 labels under both micro and macro averaging schemes. In addition, we leveraged the BERT encoder within CheXbert to obtain dense embeddings for textual similarity analysis. Each report—both ground-truth and generated—was first segmented into individual sentences using a sentence tokenizer. The BERT model was then used to encode each sentence into an embedding vector, and we computed the cosine similarity between corresponding sentences to estimate semantic alignment. The final similarity score for a report pair was obtained by averaging these sentence-level cosine similarities, following a procedure conceptually similar to that used in CXRFEScore [28].

RadGraph F1. RadGraph-based factual consistency was evaluated using the official `radgraph` library [12], accessible at https://pypi.org/project/radgraph/. We adopted the recommended RG_ER reward as the RadGraph F1 metric, which jointly measures overlap in entity and relation predictions between generated and reference reports.

CXRFEScore. We further computed CXRFEScore [28] to assess semantic and factual consistency via structured medical knowledge representations. CXRFEScore combines two components: a fact extractor and a fact encoder. We employed the publicly released CXRFEScore models (fact extractor and fact encoder) provided by the original authors. The extracted facts from both generated and ground-truth reports were encoded and compared in the resulting embedding space to produce the final factual consistency score.

RaTEScore. To assess entity-aware radiology text similarity, we utilized RaTEScore [57], available at https://pypi.org/project/RaTEScore/. Unlike standard lexical metrics, RaTEScore emphasizes crucial medical entities, such as diagnostic outcomes and anatomical details, and is designed to be robust against complex medical synonyms while remaining sensitive to negation expressions. We employed the default pipeline, which utilizes

a fine-tuned DeBERTa model for Medical Entity Recognition (NER) and BioLORD-2023-C for synonym disambiguation, to compute the alignment between generated and reference reports.

Bounding Box Metrics. For visual grounding evaluation, Intersection-over-Union (IoU) was computed using standard bounding box evaluation scripts. In cases where either the ground truth or the model output contained multiple bounding boxes for a given region or entity, we first merged all ground-truth boxes into a single region and likewise merged all predicted boxes, then computed IoU between the two resulting union regions. This avoids ambiguity when datasets provide multiple overlapping annotations. We report the mean IoU value across all evaluated samples (micro-average). Additionally, we calculate the average IoU per class and report the mean of these class-wise averages (macro-average).

9. Detailed Experimental Results

This section provides the complete, unabridged results from our experiments, including the full ablation study and per-task performance tables.

9.1. Extended Ablation Study

Table 10 provides detailed definitions for the model versions evaluated in our ablation study, and Table 11 presents comprehensive quantitative results. This section offers a step-by-step analysis of the training dynamics that led to the final CURE method. We examine the progression in four stages: the impact of baseline augmentations (v1–v2), the optimization of curriculum update schedules (v3–v5), the interaction between pre-training and learning rate scaling (v6–v11), and finally, a detailed isolation of sampling strategy effects (v12–v15).

Baseline and Data Augmentation (v1–v2). Our baseline model (**v1**) employs uniform sampling without augmentation. As observed in Table 11, introducing bounding-box-aware augmentation (**v2**) results in consistent, though modest, improvements in Phrase Grounding (PG) metrics across datasets (e.g., MS-CXR IoU improves from 0.388 to 0.398). This suggests that spatial transformations help the model generalize better to anatomical coordinates that may differ slightly from the training prototypes, without requiring changes to the model architecture.

Curriculum Learning Frequency (v3–v5). Configurations v3 through v5 explore the frequency of curriculum re-weighting updates. We observe that a longer accumulation window of 3000 steps (**v5**) yields comparable or slightly better performance compared to more frequent updates (1500 or 2000 steps). This indicates that the model may benefit from longer exposure to a fixed data distribution, giving more time to the model's performance on current tasks to plateau before the curriculum logic re-adjusts the sampling ratios.

Impact of Learning Rate on Pre-training (v6–v11). A pivotal finding is the interaction between pre-training and learning rate. In variants v6–v8 (low LR, 2e-5), Chest ImaGenome (CIG) pre-training yielded only marginal gains over the baseline. However, increasing the learning rate to 2e-4 (v9–v11) unlocked substantial improvements. Comparing **v8** (Low LR) to **v11** (High LR, CURE), we observe a sharp increase in AGRG IoU (from 0.486 to **0.601**) and Phrase Grounding MS-CXR IoU (from 0.495 to 0.552). This implies that a higher learning rate is necessary to effectively adapt the visual encoder to fine-grained anatomical text after the initial pre-training phase. Among these high-LR variants, the 3000-step pre-training schedule (**v11**) provided the most consistent performance across tasks, serving as our final **CURE** configuration.

Sampling Strategy Analysis (v12–v15). Finally, we investigate the impact of data mixing strategies. As detailed in Section 3.1, our framework defines "data sources" as specific dataset-task pairs (e.g., PadChest-GR (task: GRG) vs. MS-CXR (task: PG)). To facilitate the analysis, we define the three sampling approaches used in variants v12–v15 as follows:

- *Natural Sampling:* At the **Inter-level**, data sources are sampled strictly proportional to their size (heavily biasing training toward Chest ImaGenome). At the **Intra-level**, samples are drawn randomly without intervention, preserving the inherent clinical class imbalance.
- *Uniform Sampling:* At the **Inter-level**, all data sources are sampled with equal probability ($1/K$). At the **Intra-level**, samples are drawn such that each category (e.g., finding or anatomical region) has an equal probability of selection ($1/C$).
- *Curriculum Sampling:* Sampling probabilities are dynamically re-weighted based on error rates. At the **Inter-level**, this balances distinct data sources based on aggregate validation performance. At the **Intra-level**, this re-weights specific intra-dataset categories based on per-class error.

We analyze the impact of these strategies on both in-domain tasks and the zero-shot out-of-distribution (OOD) benchmark, VinDr-CXR.

- **The Risks of Natural Sampling (v15):** Variant **v15** employs a fully Natural strategy. While this achieves the absolute highest performance on the dominant Chest ImaGenome dataset (AGRG IoU **0.639**), it underperforms

Table 10. **Experimental Configuration Summary.** Detailed definitions of the model configurations (v1–v15) evaluated in the ablation study (see Table 11). The table outlines the progression from the baseline model to the proposed CURE method, detailing variations in data augmentation (Aug), curriculum learning (CL) schedules, Chest ImaGenome (CIG) pre-training duration, learning rates, and sampling strategies (Inter/Intra-dataset).

Model Configuration	Description
— Baseline & Augmentation —	
v1: Base (w/o Aug, w/o CL, w/o CIG, lr=2e-5)	**Baseline:** Basic multi-task fine-tuning using uniform sampling across all datasets (inter-dataset) and within datasets (intra-dataset). No data augmentation or pre-training is applied. Fine-tuned with a base learning rate of 2e-5 for 6k steps.
v2: + Aug	Identical to **v1**, but enables **bounding-box-aware** augmentations (stochastic CLAHE, RandomResizedCrop, and affine transforms). Ground-truth text coordinates are dynamically updated to match spatial changes. Horizontal flipping and color distortions are disabled.
— Curriculum Learning (CL) Frequency —	
v3: + Aug + CL(1.5k)	Extends **v2** by introducing curriculum learning (CL). The sampling distribution is re-weighted based on model performance every **1500 steps**.
v4: + Aug + CL(2k)	Same as **v3**, but the curriculum re-weighting interval is increased to every **2000 steps**.
v5: + Aug + CL(3k)	Same as **v3**, but the curriculum re-weighting interval is set to every **3000 steps**. This serves as the foundational CL schedule for subsequent experiments.
— CIG Pre-training Integration (Low LR: 2e-5) —	
v6: + Aug + CIG(1k) + CL(3k)	Introduces a pre-training phase on the Chest ImaGenome (CIG) dataset for **1000 steps** (lr=2e-5) before initializing the multi-task fine-tuning configuration of **v5**.
v7: + Aug + CIG(2k) + CL(3k)	Extends the CIG pre-training phase to **2000 steps** (lr=2e-5) before fine-tuning.
v8: + Aug + CIG(3k) + CL(3k)	Extends the CIG pre-training phase to **3000 steps** (lr=2e-5) before fine-tuning.
— Learning Rate Scaling (High LR: 2e-4) —	
v9: + Aug + CIG(1k) + CL(3k) + lr=2e-4	Replicates the structure of **v6** (1k pre-train), but significantly increases the learning rate to **2e-4** for both pre-training and fine-tuning stages.
v10: + Aug + CIG(2k) + CL(3k) + lr=2e-4	Replicates the structure of **v7** (2k pre-train) with the higher learning rate of **2e-4**.
v11 (CURE): + Aug + CIG(3k) + CL(3k) + lr=2e-4	**Proposed Method (CURE):** Replicates **v8** (3k pre-train) with the higher learning rate (2e-4). Combines prolonged pre-training, high learning rate, and 3k-step curriculum updates.
— Sampling Strategy Ablations (Based on v11) —	
v12: + Aug + CIG(3k) + Uni(Inter)/Nat(Intra) + lr=2e-4	Modification of **v11** that removes Curriculum Learning entirely. Uses **Uniform** sampling between datasets (Inter) and **Natural** distribution sampling within datasets (Intra).
v13: + Aug + CIG(3k) + CL(Inter,3k)/Nat(Intra) + lr=2e-4	Modification of **v11** that applies Curriculum Learning (3k re-weighting) **only** to the inter-dataset sampling ratios, while maintaining a **Natural** distribution for intra-dataset sampling.
v14: + Aug + CIG(3k) + Uni(Inter)/CL(Intra,3k) + lr=2e-4	Modification of **v11** that applies **Uniform** sampling between datasets (Inter), while applying **Curriculum Learning** (3k re-weighting) exclusively to intra-dataset sampling.
v15: + Aug + CIG(3k) + Nat(Inter)/Nat(Intra) + lr=2e-4	Modification of **v11** using a fully **Natural** sampling strategy (proportional to dataset size) for both inter-dataset and intra-dataset distributions.

Table 11. **Extended Ablation Study.** Performance of all ablation variants across three tasks: **AGRG** (Anatomy-Grounded Report Generation on Chest ImaGenome), **GRG** (Grounded Report Generation on PadChest-GR and VinDr-CXR), and **PG** (Phrase Grounding on MS-CXR, PadChest-GR, and VinDr-CXR). Each block reports mean Intersection-over-Union (**IoU**, micro average), CheXbert F1 (micro average, **F1**), and CXRFEScore (**CXS**) metrics. Rows v1–v5 analyze the effects of data augmentation and curriculum learning (CL); v6–v8 add CIG pre-training with a low learning rate (2e–5); v9–v11 repeat those with a higher learning rate (2e–4); and v12–v15 further explore inter- and intra-dataset sampling strategies. Best results in each column are shown in **bold**, and second-best are underlined.

Dataset abbreviations: CIG = Chest ImaGenome, PC = PadChest-GR, VD = VinDr-CXR, MS = MS-CXR.

Model Configuration	AGRG (CIG)			GRG (PC)			GRG (VD)			PG (IoU ↑)		
	IoU ↑	F1 ↑	CXS ↑	IoU ↑	F1 ↑	CXS ↑	IoU ↑	F1 ↑	CXS ↑	MS	PC	VD
MAIRA-2 (External Baseline)	0.249 ± 0.008	0.377 ± 0.016	0.357 ± 0.010	0.256 ± 0.011	**0.591** ± 0.015	**0.616** ± 0.011	0.217 ± 0.007	0.546 ± 0.008	0.591 ± 0.005	0.495 ± 0.016	0.280 ± 0.008	0.161 ± 0.005
v1: Base (w/o Aug, w/o CL, w/o CIG, lr=2e-5)	0.380 ± 0.008	0.517 ± 0.017	0.517 ± 0.011	0.171 ± 0.009	0.557 ± 0.015	0.589 ± 0.011	0.207 ± 0.006	0.586 ± 0.008	0.630 ± 0.007	0.388 ± 0.017	0.356 ± 0.007	0.191 ± 0.004
v2: + Aug	0.360 ± 0.009	0.500 ± 0.018	0.522 ± 0.011	0.185 ± 0.010	0.564 ± 0.015	0.599 ± 0.011	0.221 ± 0.007	0.614 ± 0.008	0.648 ± 0.007	0.398 ± 0.019	0.366 ± 0.007	0.203 ± 0.005
v3: + Aug + CL(1.5k)	0.399 ± 0.009	0.487 ± 0.018	0.521 ± 0.011	0.179 ± 0.010	0.564 ± 0.016	0.592 ± 0.011	0.224 ± 0.007	0.605 ± 0.008	0.630 ± 0.007	0.409 ± 0.019	0.383 ± 0.007	0.210 ± 0.005
v4: + Aug + CL(2k)	0.394 ± 0.009	0.504 ± 0.017	0.513 ± 0.011	0.193 ± 0.010	0.568 ± 0.015	0.596 ± 0.011	0.222 ± 0.006	0.611 ± 0.008	0.651 ± 0.007	0.393 ± 0.017	0.383 ± 0.007	0.196 ± 0.005
v5: + Aug + CL(3k)	0.411 ± 0.009	0.493 ± 0.017	0.526 ± 0.011	0.180 ± 0.010	0.578 ± 0.014	0.595 ± 0.012	0.217 ± 0.007	**0.626** ± 0.008	**0.671** ± 0.007	0.430 ± 0.018	0.393 ± 0.007	0.205 ± 0.005
v6: + Aug + CIG(1k) + CL(3k)	0.454 ± 0.008	0.512 ± 0.017	0.518 ± 0.011	0.195 ± 0.009	0.553 ± 0.015	0.591 ± 0.011	0.232 ± 0.006	0.582 ± 0.008	0.628 ± 0.007	0.457 ± 0.016	0.394 ± 0.007	0.219 ± 0.005
v7: + Aug + CIG(2k) + CL(3k)	0.448 ± 0.008	0.532 ± 0.017	0.533 ± 0.011	0.203 ± 0.010	0.535 ± 0.015	0.586 ± 0.011	0.227 ± 0.006	0.553 ± 0.008	0.590 ± 0.007	0.467 ± 0.016	0.403 ± 0.007	0.222 ± 0.005
v8: + Aug + CIG(3k) + CL(3k)	0.486 ± 0.008	0.530 ± 0.017	0.521 ± 0.011	0.207 ± 0.010	0.552 ± 0.015	0.582 ± 0.011	0.233 ± 0.006	0.568 ± 0.008	0.601 ± 0.007	0.495 ± 0.016	0.421 ± 0.007	0.224 ± 0.005
v9: + Aug + CIG(1k) + CL(3k) + lr=2e-4	0.607 ± 0.008	0.517 ± 0.017	0.531 ± 0.011	0.253 ± 0.010	0.522 ± 0.016	0.563 ± 0.010	0.259 ± 0.007	0.491 ± 0.008	0.529 ± 0.006	0.564 ± 0.016	0.453 ± 0.007	0.247 ± 0.005
v10: + Aug + CIG(2k) + CL(3k) + lr=2e-4	0.606 ± 0.008	0.515 ± 0.017	0.535 ± 0.011	0.258 ± 0.010	0.523 ± 0.015	0.569 ± 0.010	0.262 ± 0.007	0.449 ± 0.008	0.477 ± 0.006	**0.574** ± 0.015	0.457 ± 0.007	**0.248** ± 0.005
v11 (CURE): + Aug + CIG(3k) + CL(3k) + lr=2e-4	0.601 ± 0.008	0.529 ± 0.017	0.549 ± 0.011	0.265 ± 0.011	0.507 ± 0.015	0.574 ± 0.010	0.262 ± 0.007	0.505 ± 0.008	0.540 ± 0.007	0.552 ± 0.015	0.453 ± 0.006	0.243 ± 0.005
v12: + Aug + CIG(3k) + Uni(Inter)/Nat(Intra) + lr=2e-4	0.595 ± 0.008	0.509 ± 0.017	0.526 ± 0.011	0.263 ± 0.010	0.529 ± 0.015	0.575 ± 0.010	0.264 ± 0.007	0.489 ± 0.008	0.514 ± 0.007	0.557 ± 0.015	0.457 ± 0.007	0.245 ± 0.005
v13: + Aug + CIG(3k) + CL(Inter,3k)/Nat(Intra) + lr=2e-4	0.612 ± 0.008	**0.533** ± 0.017	0.529 ± 0.011	**0.272** ± 0.010	0.503 ± 0.016	0.550 ± 0.010	**0.266** ± 0.007	0.434 ± 0.008	0.469 ± 0.006	0.567 ± 0.016	**0.464** ± 0.007	0.243 ± 0.005
v14: + Aug + CIG(3k) + Uni(Inter)/CL(Intra,3k) + lr=2e-4	0.603 ± 0.008	0.525 ± 0.017	0.547 ± 0.011	0.246 ± 0.010	0.544 ± 0.015	0.570 ± 0.010	0.262 ± 0.007	0.496 ± 0.008	0.536 ± 0.007	0.555 ± 0.015	0.456 ± 0.007	0.244 ± 0.005
v15: + Aug + CIG(3k) + Nat(Inter)/Nat(Intra) + lr=2e-4	**0.639** ± 0.008	0.530 ± 0.017	**0.554** ± 0.011	0.000 ± 0.000	0.498 ± 0.016	0.483 ± 0.011	0.000 ± 0.000	0.604 ± 0.009	0.534 ± 0.006	0.355 ± 0.019	0.221 ± 0.006	0.162 ± 0.005

significantly on all other tasks. Most notably, it suffers a complete collapse on Grounded Report Generation (GRG), dropping to **0.000** IoU for both PadChest and VinDr-CXR. This confirms that without explicit re-

balancing, the model overfits the largest data source and fails to acquire generalizable capabilities for auxiliary tasks.

- **Uniform vs. Curriculum (v12 vs. v13):** Removing CL entirely and employing Uniform inter-dataset sampling (**v12**) effectively prevents the collapse seen in v15, yielding a very strong baseline that competes closely with the curriculum variants. However, applying Curriculum Learning to the inter-dataset mix (**v13**) offers marginal but consistent gains over the Uniform baseline across several benchmarks. For instance, v13 achieves higher grounding performance on PadChest (GRG IoU **0.272** vs. 0.263) and MS-CXR (IoU **0.567** vs. 0.557). This suggests that while Uniform sampling provides a robust foundation, dynamic re-weighting can squeeze out minor performance improvements by prioritizing harder tasks.
- **Inter- vs. Intra-Dataset Dynamics (v13 vs. v14):** We further isolate the CL logic. Variant **v13** applies CL only to the *Inter-dataset* mix, while **v14** applies CL only to the *Intra-dataset* mix. Variant **v13** outperforms **v14** on Grounded Report Generation tasks (e.g., GRG-PC IoU **0.272** vs. 0.246). Empirically, the curriculum logic at the inter-dataset level, present in v13, tends to assign higher sampling probabilities to the GRG task, as previously seen in Figure 4, compared to uniform or intra-only strategies, likely due to the higher difficulty of generating full grounded reports. While the differences are not dramatic, the results indicate that macro-level balancing between distinct data sources (Inter-CL) is a more effective driver of robustness than fine-grained category re-weighting (Intra-CL).

Ultimately, while Uniform sampling (**v12**) proves to be a highly effective strategy for multi-task stability, the Curriculum-based methods (v11/v13) demonstrate the capacity to further refine performance on challenging tasks like Grounded Report Generation without compromising the baseline capabilities.

9.2. Sensitivity to α (Weighting Term)

As detailed in Section 6.3, our curriculum protocol computes an aggregate performance score s_i for each data source and/or intra-dataset class using a weighted average of localization (IoU) and semantic alignment (CXRFEScore). This balance is governed by the parameter α:

$$s_i = \alpha \cdot \text{IoU}_i + (1 - \alpha) \cdot \text{CXRFEScore}_i \quad (4)$$

In our primary experiments, we set $\alpha = 0.8$ to explicitly prioritize improvements in spatial localization accuracy. To empirically understand the sensitivity of the CURE framework to this parameter, we conducted an ablation study varying $\alpha \in \{0.0, 0.25, 0.5, 0.75, 0.8, 1.0\}$.

Due to the computational cost of full training runs, these specific ablation experiments were restricted to 3000 steps of training in the AGRG task on the Chest ImaGenome dataset. The results, including bootstrapped standard deviations, are presented in Table 12.

Table 12. **Sensitivity Analysis of the Curriculum Weighting Term (α).** Performance metrics on the Chest ImaGenome dataset (AGRG task) after 3000 training steps across different values of α. Higher values of α heavily weight the IoU metric during curriculum updates, while lower values prioritize the text-based semantic metric (CXRFEScore). We report mean Intersection-over-Union (**IoU**, ↑), CheXbert F1 (Micro/Macro averages, ↑), CheXbert cosine similarity (**Cos.**, ↑), and CXRFEScore (**CXS**, ↑). **Bold** indicates the best result per column.

Method	IoU ↑	F1-Mi ↑	F1-Ma ↑	Cos. ↑	CXS ↑
CURE ($\alpha = 0.0$)	0.582 ± 0.008	0.527 ± 0.017	0.226 ± 0.017	0.688 ± 0.009	0.556 ± 0.011
CURE ($\alpha = 0.25$)	0.591 ± 0.008	0.522 ± 0.018	0.224 ± 0.019	0.692 ± 0.008	0.538 ± 0.012
CURE ($\alpha = 0.5$)	0.599 ± 0.008	0.529 ± 0.017	$\mathbf{0.250} \pm 0.017$	$\mathbf{0.697} \pm 0.009$	0.541 ± 0.012
CURE ($\alpha = 0.75$)	0.603 ± 0.008	0.523 ± 0.017	0.228 ± 0.016	0.690 ± 0.009	0.545 ± 0.012
CURE ($\alpha = 0.8$)	$\mathbf{0.616} \pm 0.008$	0.519 ± 0.018	0.215 ± 0.016	0.686 ± 0.009	0.545 ± 0.011
CURE ($\alpha = 1.0$)	0.608 ± 0.008	$\mathbf{0.533} \pm 0.017$	0.236 ± 0.017	0.695 ± 0.009	$\mathbf{0.558} \pm 0.011$

As anticipated, increasing α generally yields consistent improvements in visual grounding. The model achieves its peak spatial accuracy at $\alpha = 0.8$ (IoU **0.616**), confirming our design choice to prioritize visual grounding accuracy during curriculum updates.

However, decreasing α toward 0.0 (which theoretically forces the curriculum to prioritize text generation quality) does not yield a monotonic improvement in clinical text metrics like F1-Macro or CXRFEScore. Instead, text performance fluctuates, peaking at moderate values ($\alpha = 0.5$ or $\alpha = 1.0$).

We hypothesize that this behavior is a limitation of the current intra-dataset balancing strategy. As noted in Section 3.2, the intra-dataset curriculum for AGRG is designed to balance exposure across *anatomical locations*, but it does not actively re-weight the distribution of *clinical findings*. Consequently, even when a lower α signals the model to focus heavily on the text generation objective, the model's ability to improve its text metrics is constrained by the natural, long-tailed imbalance of pathological findings inherent to the dataset. Future iterations of the CURE framework could address this by implementing finer-grained, multi-dimensional balancing strategies that stratify samples based simultaneously on both anatomical regions and the prevalence of specific clinical findings.

9.3. Full Results Tables

The following tables contain the complete, unabridged results for each evaluation task and dataset, from which the summary tables in the main paper were derived.

9.3.1. Anatomy-Grounded Report Generation (AGRG)

Table 13 presents a comprehensive breakdown of performance on the Anatomy-Grounded Report Generation

(AGRG) task, isolating the effects of pre-training strategies and multi-task fine-tuning configurations.

Trade-offs Between Grounding and Clinical Label Accuracy. Comparing the baseline (**v1**) with the optimized variants (v11–v15) highlights a stark contrast in the rate of improvement between spatial and text-based metrics. While the optimized models achieve substantial gains in spatial precision (with IoU jumping from 0.380 to > 0.59), the text generation metrics do not scale proportionally. Semantic metrics such as CheXbert Cosine Similarity and CXR-FEScore show only modest improvements, and the baseline model actually retains the highest CheXbert F1-Macro score (0.251), whereas most high-IoU variants fluctuate between 0.21 and 0.25. We hypothesize that this discrepancy stems from our current curriculum design. For the AGRG task, the intra-dataset re-weighting strategy is explicitly designed to balance *anatomical locations* to ensure robust localization, but it does not currently account for the highly imbalanced distribution of *clinical findings*. Consequently, while we observe dramatic improvements in spatial grounding and stable performance on dominant text classes (F1-Micro), the model does not fully benefit from re-balancing rare pathological conditions. This highlights a clear avenue for future work: designing a multidimensional re-weighting strategy that simultaneously targets anatomical diversity and the distribution of rare clinical findings to improve semantic report quality on long-tailed conditions.

Pre-training Efficiency. The "Chest ImaGenome (CIG) Pre-training Only" block highlights the impact of learning rate scaling. With a conservative learning rate (2e-5), extending pre-training from 1000 to 3000 steps yields only marginal IoU gains (0.378 $\rightarrow$ 0.430). Conversely, increasing the learning rate to 2e-4 results in a substantial improvement, with the 3000-step high-LR variant achieving an IoU of 0.596 even before multi-task fine-tuning. This confirms that aligning visual features with fine-grained anatomy-grounded reports benefits from more aggressive optimization during the initial training stages.

Performance of Sampling Strategies. Among the final sampling variants, we observe that **v15** (Natural Sampling) achieves the highest scores across all metrics in this specific task (IoU **0.639**, Cos. Sim. **0.694**, CXS **0.554**). This result is expected given the data distribution: **v15** undergoes 3000 steps of pre-training and 6000 steps of fine-tuning where samples are drawn proportional to dataset size. Since Chest ImaGenome dominates the training mixture, **v15** is effectively trained on AGRG for the vast majority of these $\sim$9,000 steps. However, as detailed in the ablation study (Section 9.1), this specialization leads to severe degradation on complementary tasks (GRG and PG) on other datasets. The strategies that actively intervene on the data distribution—Uniform (**v12**) and the Curriculum variants (**v11**, **v13**, **v14**)—maintain competitive in-domain performance (IoU $\sim$0.60, CXS $\sim$0.53–0.55) while preventing the task collapse observed in **v15**. Notably, all proposed variants (v9–v15) significantly outperform the external MAIRA-2 baseline in both spatial grounding (IoU $\sim$0.59–0.64 vs. 0.249) and semantic alignment (Cos. Sim. > 0.67 and CXS > 0.52 vs. 0.662 and 0.467, respectively).

9.3.2. Phrase Grounding (PG)

Table 14 details the Phrase Grounding performance across three diverse benchmarks: MS-CXR (in-domain), PadChest-GR (in-domain), and VinDr-CXR (zero-shot, unseen distribution).

Generalization via Augmentation. Adding data augmentation (**v2**) to the baseline (**v1**) yields modest but consistent spatial improvements. For instance, MS-CXR IoU Micro increases from 0.388 to 0.398, and zero-shot VinDr-CXR IoU improves from 0.191 to 0.203. This indicates that bounding-box-aware augmentations effectively reduce overfitting and help the model generalize to varied image acquisitions.

Progressive Improvements: Curriculum, Pre-training, and Learning Rate. The results demonstrate a cumulative benefit from each component of the CURE pipeline. First, introducing curriculum learning alone (variants **v3–v5**) yields a moderate gain over the augmented baseline (e.g., **v5** reaches 0.430 IoU on MS-CXR vs. 0.398 for **v2**). Second, adding Chest ImaGenome pre-training with a conservative learning rate (variants **v6–v8**) pushes performance further, with **v8** reaching 0.495 IoU. Finally, the most dramatic jump occurs when increasing the learning rate to 2e-4 for *both* the pre-training and multi-task fine-tuning phases (variants **v9–v11**). Comparing **v8** (Low LR) to **v11** (High LR), we observe an improvement of over 5 points on MS-CXR (0.495 $\rightarrow$ 0.552). This confirms that a higher learning rate is essential throughout the entire pipeline to fully align visual features with text and escape local minima.

Sampling Strategy Dynamics. Analyzing the sampling strategies (v10–v15) reveals distinct performance profiles across datasets:

- **Natural Sampling Failure:** Variant **v15** suffers a severe regression, dropping to 0.355 IoU on MS-CXR (worse than the un-augmented baseline **v1**). This confirms that without intervention, the dominance of AGRG data overwhelms the signal from smaller grounding datasets.
- **Peak Performance (v10 vs. v13):** While the proposed CURE model (**v11**) is highly competitive, the absolute peak performance for phrase grounding is split between

Table 13. **Detailed Results for Anatomy-Grounded Report Generation (AGRG).** Performance of baseline models, pre-training-only checkpoints, and the full set of multi-task fine-tuning ablation variants (v1–v15) on the Chest ImaGenome test subset. We report mean Intersection-over-Union (**IoU**, ↑), CheXbert F1 (Micro/Macro averages, ↑), CheXbert cosine similarity (**Cos.**, ↑), and CXRFEScore (**CXS**, ↑). **Bold** indicates the best result per column; underlined indicates the second best.

Model Variant	**IoU** ↑	**F1-Mi** ↑	**F1-Ma** ↑	**Cos.** ↑	**CXS** ↑
— Baselines —					
MAIRA-2	0.249 ± 0.008	0.377 ± 0.016	0.098 ± 0.009	0.587 ± 0.010	0.357 ± 0.010
MedGemma-4B-IT	–	0.266 ± 0.012	0.227 ± 0.014	0.662 ± 0.004	0.467 ± 0.006
— Chest ImaGenome (CIG) Pre-training Only —					
CIG Pre-train (1k steps, lr=2e-5)	0.378 ± 0.008	0.530 ± 0.017	0.249 ± 0.023	0.670 ± 0.009	0.510 ± 0.012
CIG Pre-train (2k steps, lr=2e-5)	0.402 ± 0.008	0.528 ± 0.016	0.215 ± 0.015	0.666 ± 0.009	0.512 ± 0.011
CIG Pre-train (3k steps, lr=2e-5)	0.430 ± 0.008	0.526 ± 0.017	0.221 ± 0.015	0.672 ± 0.009	0.513 ± 0.012
CIG Pre-train (1k steps, lr=2e-4)	0.501 ± 0.008	0.525 ± 0.017	0.237 ± 0.017	0.675 ± 0.009	0.545 ± 0.011
CIG Pre-train (2k steps, lr=2e-4)	0.590 ± 0.008	0.525 ± 0.017	0.242 ± 0.018	0.686 ± 0.009	0.544 ± 0.011
CIG Pre-train (3k steps, lr=2e-4)	0.596 ± 0.008	0.524 ± 0.017	0.235 ± 0.017	0.688 ± 0.009	0.551 ± 0.011
— Multi-task Fine-tuning Variants (v1–v15) —					
v1: Base (w/o Aug, w/o CL, w/o CIG, lr=2e-5)	0.380 ± 0.008	0.517 ± 0.017	**0.251** ± 0.017	0.665 ± 0.009	0.517 ± 0.011
v2: + Aug	0.360 ± 0.009	0.500 ± 0.018	0.211 ± 0.015	0.660 ± 0.009	0.522 ± 0.011
v3: + Aug + CL(1.5k)	0.399 ± 0.009	0.487 ± 0.018	0.207 ± 0.015	0.653 ± 0.009	0.521 ± 0.011
v4: + Aug + CL(2k)	0.394 ± 0.009	0.504 ± 0.017	0.212 ± 0.015	0.660 ± 0.009	0.513 ± 0.011
v5: + Aug + CL(3k)	0.411 ± 0.009	0.493 ± 0.017	0.197 ± 0.011	0.655 ± 0.009	0.526 ± 0.011
v6: + Aug + CIG(1k) + CL(3k)	0.454 ± 0.008	0.512 ± 0.017	0.213 ± 0.016	0.656 ± 0.009	0.518 ± 0.011
v7: + Aug + CIG(2k) + CL(3k)	0.448 ± 0.008	0.532 ± 0.017	**0.251** ± 0.021	0.673 ± 0.009	0.533 ± 0.011
v8: + Aug + CIG(3k) + CL(3k)	0.486 ± 0.008	0.530 ± 0.017	0.226 ± 0.016	0.671 ± 0.009	0.521 ± 0.011
v9: + Aug + CIG(1k) + CL(3k) + lr=2e-4	0.607 ± 0.008	0.517 ± 0.017	0.223 ± 0.016	0.679 ± 0.009	0.531 ± 0.011
v10: + Aug + CIG(2k) + CL(3k) + lr=2e-4	0.606 ± 0.008	0.515 ± 0.017	0.212 ± 0.014	0.675 ± 0.009	0.535 ± 0.011
v11 (CURE): + Aug + CIG(3k) + CL(3k) + lr=2e-4	0.601 ± 0.008	0.529 ± 0.017	0.234 ± 0.018	0.691 ± 0.009	0.549 ± 0.011
v12: + Aug + CIG(3k) + Uni(Inter)/Nat(Intra) + lr=2e-4	0.595 ± 0.008	0.509 ± 0.017	0.218 ± 0.020	0.674 ± 0.009	0.526 ± 0.011
v13: + Aug + CIG(3k) + CL(Inter,3k)/Nat(Intra) + lr=2e-4	0.612 ± 0.008	**0.533** ± 0.017	0.249 ± 0.021	0.687 ± 0.009	0.529 ± 0.011
v14: + Aug + CIG(3k) + Uni(Inter)/CL(Intra,3k) + lr=2e-4	0.603 ± 0.008	0.525 ± 0.017	0.227 ± 0.018	0.685 ± 0.009	0.547 ± 0.011
v15: + Aug + CIG(3k) + Nat(Inter)/Nat(Intra) + lr=2e-4	**0.639** ± 0.008	0.530 ± 0.017	0.250 ± 0.018	**0.694** ± 0.009	**0.554** ± 0.011

Table 14. **Detailed Results for Phrase Grounding (PG).** We report Micro-Average IoU (**IoU Mi.** ↑) and Macro-Average IoU (**IoU Ma.** ↑) on three test sets: MS-CXR, PadChest-GR, and zero-shot VinDr-CXR. **Bold** indicates best; underlined indicates second best.

Model Variant	**MS-CXR**		**PadChest-GR**		**VinDr-CXR (Zero-Shot)**	
	IoU Mi. ↑	**IoU Ma.** ↑	**IoU Mi.** ↑	**IoU Ma.** ↑	**IoU Mi.** ↑	**IoU Ma.** ↑
MAIRA-2	0.495 ± 0.016	0.453 ± 0.016	0.280 ± 0.008	0.288 ± 0.009	0.161 ± 0.005	0.114 ± 0.010
v1: Base (w/o Aug, w/o CL, w/o CIG, lr=2e-5)	0.388 ± 0.017	0.344 ± 0.016	0.356 ± 0.007	0.345 ± 0.007	0.191 ± 0.004	0.144 ± 0.010
v2: + Aug	0.398 ± 0.019	0.353 ± 0.016	0.366 ± 0.007	0.360 ± 0.007	0.203 ± 0.005	0.153 ± 0.007
v3: + Aug + CL(1.5k)	0.409 ± 0.019	0.369 ± 0.016	0.383 ± 0.007	0.382 ± 0.007	0.210 ± 0.005	0.154 ± 0.007
v4: + Aug + CL(2k)	0.393 ± 0.017	0.348 ± 0.015	0.383 ± 0.007	0.381 ± 0.008	0.196 ± 0.005	0.145 ± 0.007
v5: + Aug + CL(3k)	0.430 ± 0.018	0.377 ± 0.016	0.393 ± 0.007	0.397 ± 0.008	0.205 ± 0.005	0.155 ± 0.006
v6: + Aug + CIG(1k) + CL(3k)	0.457 ± 0.016	0.405 ± 0.014	0.394 ± 0.007	0.391 ± 0.008	0.219 ± 0.005	0.167 ± 0.011
v7: + Aug + CIG(2k) + CL(3k)	0.467 ± 0.016	0.428 ± 0.015	0.403 ± 0.007	0.399 ± 0.008	0.222 ± 0.005	0.160 ± 0.006
v8: + Aug + CIG(3k) + CL(3k)	0.495 ± 0.016	0.446 ± 0.015	0.421 ± 0.007	0.419 ± 0.008	0.224 ± 0.005	0.173 ± 0.011
v9: + Aug + CIG(1k) + CL(3k) + lr=2e-4	0.564 ± 0.016	0.514 ± 0.016	0.453 ± 0.007	0.443 ± 0.007	0.247 ± 0.005	0.203 ± 0.011
v10: + Aug + CIG(2k) + CL(3k) + lr=2e-4	**0.574** ± 0.015	**0.526** ± 0.014	0.457 ± 0.007	0.445 ± 0.008	**0.248** ± 0.005	**0.206** ± 0.011
v11 (CURE): + Aug + CIG(3k) + CL(3k) + lr=2e-4	0.552 ± 0.015	0.495 ± 0.015	0.453 ± 0.006	0.438 ± 0.007	0.243 ± 0.005	0.205 ± 0.012
v12: + Aug + CIG(3k) + Uni(Inter)/Nat(Intra) + lr=2e-4	0.557 ± 0.015	0.507 ± 0.014	0.457 ± 0.007	0.448 ± 0.007	0.245 ± 0.005	0.198 ± 0.007
v13: + Aug + CIG(3k) + CL(Inter,3k)/Nat(Intra) + lr=2e-4	0.567 ± 0.016	0.515 ± 0.014	**0.464** ± 0.007	**0.455** ± 0.007	0.243 ± 0.005	0.199 ± 0.011
v14: + Aug + CIG(3k) + Uni(Inter)/CL(Intra,3k) + lr=2e-4	0.555 ± 0.015	0.496 ± 0.013	0.456 ± 0.007	0.451 ± 0.008	0.244 ± 0.005	0.203 ± 0.013
v15: + Aug + CIG(3k) + Nat(Inter)/Nat(Intra) + lr=2e-4	0.355 ± 0.019	0.277 ± 0.016	0.221 ± 0.006	0.206 ± 0.005	0.162 ± 0.005	0.099 ± 0.004

v10 and **v13**. Variant **v10** (2k pre-training) achieves the highest scores on MS-CXR (**0.574**) and the zero-shot VinDr-CXR (**0.248**), suggesting that a slightly shorter pre-training phase may occasionally favor pure localization tasks. Conversely, **v13** (Inter-CL) achieves the best performance on PadChest-GR (**0.464**), indicating that dynamic inter-dataset balancing effectively captures the nuances of that specific distribution.

- **Overall Robustness:** Despite these minor variations, all high-LR curriculum variants (v10, v11, v13, v14) significantly outperform the external MAIRA-2 baseline (e.g., $\sim$0.24 vs. 0.16 on VinDr-CXR), validating the general effectiveness of the proposed framework.

9.3.3. Grounded Report Generation (GRG)

Table 15 presents the ablation results for the Grounded Report Generation (GRG) task on PadChest-GR and the zero-shot VinDr-CXR benchmark. This task is the most challenging in our suite, requiring the model to simultaneously generate a full radiology report and localize every mentioned finding.

Localization vs. Clinical Metrics. A clear divergence emerges when comparing the low-learning-rate variants (**v1–v8**) with the high-learning-rate variants (**v9–v14**). The low-LR models often achieve higher scores on clinical text metrics; for instance, **v5** achieves a zero-shot F1-Micro of **0.626** on VinDr-CXR but has limited localization accuracy (IoU 0.217). Conversely, the high-LR variants significantly boost visual grounding (e.g., **v13** reaches IoU **0.266**) but often see a regression in text metrics (F1-Micro drops to 0.434). This pattern suggests that while aggressive updates are necessary to learn the structural constraints of the GRG task (i.e., outputting bounding box coordinates), our current training protocol heavily favors the grounding objective. Future work is likely needed to design more sophisticated re-balancing strategies—such as balancing positive versus negative findings or specific finding classes—to simultaneously enhance clinical reporting metrics without sacrificing localization performance.

Benchmarking Against MAIRA-2. The PadChest-GR results are particularly significant. This task theoretically favors the MAIRA-2 baseline, which benefits from training on both PadChest-GR and the large, proprietary USMix dataset [3] (containing $\sim$70k grounded reports). Despite this data disadvantage, our high-LR variants (**v10**, **v11**, **v12**, **v13**) consistently surpass MAIRA-2 in localization performance (e.g., **v13** IoU **0.272** vs. 0.256). On the zero-shot VinDr-CXR benchmark, this trend holds, with most high-LR variants significantly outperforming MAIRA-2 (**0.266** vs. 0.217). However, **v15** (Natural Sampling) illustrates a critical failure mode: it achieves an IoU of **0.000** on both datasets yet records the highest Cosine Similarity on VinDr-CXR (**0.858**). This suggests that due to the overwhelming dominance of AGRG data (Chest ImaGenome), the model fails to learn the specific formatting requirements of the minority GRG task. Despite this, when given the GRG instruction to "Generate a grounded report", it tends to behave like a standard captioning model: it generates clinically plausible text, but these descriptions are heavily biased by the style of the mini-reports used in the AGRG task. The task collapse here is thus a failure to acquire the correct output format due to extreme data imbalance, rather than a total loss of visual understanding.

Impact of Sampling Strategies. We focus on the comparison between **v12** (Uniform Inter-sampling) and **v13** (Curriculum Inter-sampling). As noted in Section 3.1, PadChest-GR does not utilize intra-dataset curriculum reweighting; therefore, differences in performance stem primarily from how the model balances the distinct data sources. Variant **v13** achieves the highest IoU on both datasets, surpassing the Uniform baseline (**v12**). This indicates that dynamically up-weighting the GRG data sources (based on error rates) helps the model prioritize the visual grounding objective more effectively than static uniform sampling. While the margins are modest, **v13** consistently provides the most robust grounding performance across the evaluated benchmarks.

9.3.4. Standard Report Generation (MIMIC-CXR)

Table 16 provides a detailed breakdown of report generation performance on the MIMIC-CXR test set. We analyze how different inference protocols ranging from single-prompt generation to multi-location concatenation affect the trade-off between precision, recall, and semantic alignment.

Impact of Anatomical Granularity (N). A unique feature of the CURE framework is the ability to modulate the "resolution" of the generated report by varying the number of queried anatomical locations (N). The compositions of these sets are detailed in Table 17.

- **AGRG-9 (High Precision):** By querying only 9 core locations, the model achieves high precision (P-Mi: **0.639**), comparable to the grounded reports of MAIRA-2 (P-Mi: **0.639**). This configuration also ties for the highest RadGraph F1 score among the CURE variants, making it comparatively more suitable for scenarios where minimizing false positives is a priority.
- **AGRG-29 (Balanced Supervision):** This set comprises the 29 anatomical locations for which Chest ImaGenome provides both bounding box and text supervision. Using this configuration yields a strong balance of metrics, notably achieving the second-highest RaTEScore (<u>0.592</u>), validating the quality of training on fully grounded data.
- **AGRG-38 (High Recall):** Expanding to 38 locations includes peripheral areas (e.g., neck, chest wall) that possess text supervision but lack bounding box annotations in the training data. Forcing the model to scrutinize these areas results in the highest Recall (R-Mi 0.770) and CheXbert Cosine Similarity (<u>0.793</u>) among the purely anatomical approaches. However, this exhaustive search introduces a trade-off: while recall improves, precision

Table 15. **Detailed Results for Grounded Report Generation (GRG).** Performance on PadChest-GR and the zero-shot VinDr-CXR benchmark. We report mean IoU (↑), CheXbert F1 (micro/macro, ↑), CheXbert cosine similarity (**Cos.**, ↑), and CXRFEScore (**CXS**, ↑). High-learning-rate variants (v9–v13) consistently achieve superior localization (IoU) compared to baselines. **Bold** indicates best; underlined indicates second best.

Model Variant	PadChest-GR					VinDr-CXR (Zero-Shot)				
	IoU ↑	F1-Mi ↑	F1-Ma ↑	Cos. ↑	CXS ↑	IoU ↑	F1-Mi ↑	F1-Ma ↑	Cos. ↑	CXS ↑
— Baselines —										
MAIRA-2	0.256 ± 0.011	**0.591** ± 0.015	**0.321** ± 0.019	**0.844** ± 0.004	**0.616** ± 0.011	0.217 ± 0.007	0.546 ± 0.008	0.256 ± 0.011	0.824 ± 0.002	0.591 ± 0.005
MedGemma-4B-IT	–	0.144 ± 0.009	0.203 ± 0.014	0.733 ± 0.003	0.517 ± 0.005	–	0.209 ± 0.006	0.212 ± 0.008	0.779 ± 0.001	0.596 ± 0.003
— Multi-task Fine-tuning Variants (v1–v15) —										
v1: Base (w/o Aug, w/o CL, w/o CIG, lr=2e-5)	0.171 ± 0.009	0.557 ± 0.015	0.246 ± 0.012	0.829 ± 0.005	0.589 ± 0.011	0.207 ± 0.006	0.586 ± 0.008	0.247 ± 0.012	0.835 ± 0.003	0.630 ± 0.007
v2: + Aug	0.185 ± 0.010	0.564 ± 0.015	0.246 ± 0.012	0.839 ± 0.005	0.599 ± 0.011	0.221 ± 0.007	0.614 ± 0.008	0.277 ± 0.011	0.834 ± 0.003	0.648 ± 0.007
v3: + Aug + CL(1.5k)	0.179 ± 0.010	0.564 ± 0.016	0.239 ± 0.011	0.835 ± 0.005	0.592 ± 0.011	0.224 ± 0.007	0.605 ± 0.008	0.251 ± 0.011	0.816 ± 0.003	0.630 ± 0.007
v4: + Aug + CL(2k)	0.193 ± 0.010	0.568 ± 0.015	0.246 ± 0.012	0.838 ± 0.005	0.596 ± 0.011	0.222 ± 0.006	0.611 ± 0.008	0.247 ± 0.010	0.842 ± 0.003	0.651 ± 0.007
v5: + Aug + CL(3k)	0.180 ± 0.010	0.578 ± 0.014	0.256 ± 0.010	0.843 ± 0.005	0.595 ± 0.012	0.217 ± 0.007	**0.626** ± 0.008	0.253 ± 0.009	0.843 ± 0.003	**0.671** ± 0.007
v6: + Aug + CIG(1k) + CL(3k)	0.195 ± 0.009	0.553 ± 0.015	0.279 ± 0.016	0.837 ± 0.005	0.591 ± 0.011	0.232 ± 0.006	0.582 ± 0.008	**0.280** ± 0.013	0.838 ± 0.003	0.628 ± 0.007
v7: + Aug + CIG(2k) + CL(3k)	0.203 ± 0.010	0.535 ± 0.015	0.277 ± 0.021	0.829 ± 0.005	0.586 ± 0.011	0.227 ± 0.006	0.553 ± 0.008	0.252 ± 0.009	0.828 ± 0.003	0.590 ± 0.007
v8: + Aug + CIG(3k) + CL(3k)	0.207 ± 0.010	0.552 ± 0.015	0.293 ± 0.020	0.834 ± 0.005	0.582 ± 0.011	0.233 ± 0.006	0.568 ± 0.008	0.248 ± 0.009	0.841 ± 0.003	0.601 ± 0.007
v9: + Aug + CIG(1k) + CL(3k) + lr=2e-4	0.253 ± 0.010	0.522 ± 0.016	0.313 ± 0.026	0.820 ± 0.004	0.563 ± 0.010	0.259 ± 0.007	0.491 ± 0.008	0.235 ± 0.009	0.814 ± 0.003	0.529 ± 0.006
v10: + Aug + CIG(2k) + CL(3k) + lr=2e-4	0.258 ± 0.010	0.523 ± 0.015	0.265 ± 0.013	0.820 ± 0.005	0.569 ± 0.010	0.262 ± 0.007	0.449 ± 0.008	0.236 ± 0.008	0.799 ± 0.003	0.477 ± 0.006
v11 (CURE): + Aug + CIG(3k) + CL(3k) + lr=2e-4	0.265 ± 0.011	0.507 ± 0.015	0.270 ± 0.013	0.819 ± 0.005	0.574 ± 0.010	0.262 ± 0.007	0.505 ± 0.008	0.246 ± 0.009	0.832 ± 0.003	0.540 ± 0.007
v12: + Aug + CIG(3k) + Uni(Inter)/Nat(Intra) + lr=2e-4	0.263 ± 0.010	0.529 ± 0.015	0.277 ± 0.014	0.823 ± 0.005	0.575 ± 0.010	0.264 ± 0.007	0.489 ± 0.008	0.244 ± 0.009	0.819 ± 0.003	0.514 ± 0.007
v13: + Aug + CIG(3k) + CL(Inter,3k)/Nat(Intra) + lr=2e-4	**0.272** ± 0.010	0.503 ± 0.016	0.280 ± 0.018	0.812 ± 0.005	0.550 ± 0.010	**0.266** ± 0.007	0.434 ± 0.008	0.231 ± 0.009	0.795 ± 0.003	0.469 ± 0.006
v14: + Aug + CIG(3k) + Uni(Inter)/CL(Intra,3k) + lr=2e-4	0.246 ± 0.010	0.544 ± 0.015	0.317 ± 0.019	0.824 ± 0.005	0.570 ± 0.010	0.262 ± 0.007	0.496 ± 0.008	0.245 ± 0.010	0.815 ± 0.003	0.536 ± 0.007
v15: + Aug + CIG(3k) + Nat(Inter)/Nat(Intra) + lr=2e-4	0.000 ± 0.000	0.498 ± 0.016	0.193 ± 0.016	0.788 ± 0.005	0.483 ± 0.011	0.000 ± 0.000	0.604 ± 0.009	0.193 ± 0.016	**0.858** ± 0.003	0.534 ± 0.006

Table 16. **Results for Report Generation (RG) on the MIMIC-CXR test set.** We evaluate state-of-the-art baselines, a model fine-tuned solely for report generation (**MedGemma-FT (RG)**), and the proposed **CURE** model. Notably, since CURE is multi-task, we explore different inference protocols: **GRG** (generating a single grounded report), **AGRG-N** (concatenating location-specific descriptions for N anatomical regions), and their combinations. We report CheXbert F1, Precision (P), and Recall (R) (Micro/Macro averaged), CheXbert Cosine Similarity (**Cos.**), CXRFEScore (**CXS**), RaTEScore (**RaTES**), and RadGraph F1 (**RadF1**). **Bold** and underlined values indicate the best and second-best scores.

Model / Inference Protocol	F1-Ma ↑	F1-Mi ↑	P-Ma ↑	P-Mi ↑	R-Ma ↑	R-Mi ↑	Cos. ↑	CXS ↑	RaTES ↑	RadF1 ↑
— Baselines —										
CXRMate-RRG24	0.414 ± 0.006	**0.589** ± 0.004	**0.493** ± 0.012	0.617 ± 0.005	0.415 ± 0.006	0.563 ± 0.005	0.764 ± 0.001	**0.656** ± 0.002	0.577 ± 0.002	**0.255** ± 0.002
MAIRA-2 (w/ grounding)	0.304 ± 0.006	0.489 ± 0.005	0.442 ± 0.021	**0.639** ± 0.006	0.283 ± 0.006	0.397 ± 0.005	0.751 ± 0.002	0.603 ± 0.002	0.496 ± 0.002	0.120 ± 0.002
MAIRA-2 (w/o grounding)	0.386 ± 0.006	0.554 ± 0.004	0.425 ± 0.009	0.578 ± 0.005	0.384 ± 0.006	0.533 ± 0.005	0.693 ± 0.002	0.576 ± 0.002	0.501 ± 0.002	0.143 ± 0.002
MedGemma-4B-IT	0.382 ± 0.004	0.547 ± 0.004	0.332 ± 0.005	0.452 ± 0.004	0.494 ± 0.005	0.692 ± 0.005	0.714 ± 0.001	0.580 ± 0.002	0.532 ± 0.001	0.112 ± 0.001
— Specialized Fine-tuning —										
MedGemma-FT (RG only)	0.353 ± 0.006	0.520 ± 0.005	0.469 ± 0.016	0.621 ± 0.006	0.323 ± 0.005	0.447 ± 0.005	0.753 ± 0.002	0.624 ± 0.002	0.536 ± 0.002	0.203 ± 0.002
— CURE Inference Strategies (Single Model) —										
CURE (GRG Prompt)	0.314 ± 0.006	0.463 ± 0.005	0.442 ± 0.009	0.605 ± 0.006	0.290 ± 0.006	0.376 ± 0.004	0.725 ± 0.002	0.526 ± 0.002	0.447 ± 0.002	0.077 ± 0.002
CURE (AGRG-9)	0.230 ± 0.004	0.443 ± 0.005	0.432 ± 0.014	**0.639** ± 0.006	0.225 ± 0.004	0.339 ± 0.004	0.762 ± 0.002	0.608 ± 0.002	0.557 ± 0.002	0.200 ± 0.002
CURE (AGRG-9 + GRG)	0.355 ± 0.006	0.528 ± 0.004	0.436 ± 0.008	0.600 ± 0.006	0.342 ± 0.006	0.472 ± 0.005	0.784 ± 0.001	0.640 ± 0.002	0.572 ± 0.002	0.200 ± 0.002
CURE (AGRG-29)	0.400 ± 0.005	0.559 ± 0.004	0.355 ± 0.008	0.446 ± 0.004	0.539 ± 0.005	0.749 ± 0.004	0.783 ± 0.001	0.645 ± 0.002	0.592 ± 0.001	0.181 ± 0.001
CURE (AGRG-29 + GRG)	**0.415** ± 0.005	0.562 ± 0.004	0.365 ± 0.010	0.439 ± 0.004	0.582 ± 0.005	0.781 ± 0.004	0.792 ± 0.001	0.655 ± 0.002	**0.597** ± 0.001	0.176 ± 0.001
CURE (AGRG-38)	0.395 ± 0.004	0.534 ± 0.004	0.354 ± 0.017	0.408 ± 0.004	0.593 ± 0.005	0.770 ± 0.004	0.793 ± 0.001	0.632 ± 0.001	0.577 ± 0.001	0.172 ± 0.001
CURE (AGRG-38 + GRG)	0.406 ± 0.004	0.536 ± 0.004	0.360 ± 0.022	0.404 ± 0.004	**0.628** ± 0.005	**0.798** ± 0.004	**0.800** ± 0.001	0.642 ± 0.001	0.583 ± 0.001	0.169 ± 0.001

drops significantly (P-Mi 0.408) compared to the concise AGRG-9 configuration (P-Mi 0.639), leading to slightly lower aggregate F1 scores compared to the AGRG-29 configuration.

Baseline Performance Analysis. We observe distinct performance profiles across the evaluated baselines. **CXRMate-RRG24** [33] establishes a strong benchmark, achieving the highest RadGraph F1 (**0.255**), CheXbert F1-Micro (**0.589**), and CXRFEScore (**0.656**), reflecting its optimization via reinforcement learning for clinical correctness. For **MAIRA-2**, the inference mode dictates a clear trade-off: the grounded mode maximizes precision (P-Mi **0.639**) but acts as a constraint that limits recall (0.397), whereas disabling grounding improves clinical finding detection (F1-Mi 0.554) but degrades semantic alignment (Cosine drops to 0.693). Finally, comparing the **MedGemma** variants reveals the impact of domain adaptation. The base **MedGemma-4B-IT** exhibits high recall (0.692) but low precision (0.452) and structural accuracy (RadF1 0.112). Surprisingly, fine-tuning solely on reports (**MedGemma-FT (RG)**) resulted in lower CheXbert F1 scores compared to the base model (e.g., F1-Ma 0.353 vs 0.382), although it significantly improved precision to 0.621 and achieved the second-best RadGraph F1 (0.203). However, this specialized baseline still lags behind the multi-task CURE variants in multiple metrics, such as RaTEScore (**0.597** vs. 0.536) and Cosine Similarity (0.792 vs. 0.753). This suggests that

Table 17. **Anatomical Query Configurations.** Definition of the location sets used for the AGRG inference protocols. **AGRG-29** represents the set of locations with complete supervision (Bounding Box + Text) in Chest ImaGenome. **AGRG-38** extends this to include locations that have only text supervision. **AGRG-9** is a subset of core locations.

Config.	Anatomical Locations Included
AGRG-9	*Core Locations:* Abdomen, Cardiac Silhouette, Left/Right Costophrenic Angle, Left/Right Lung, Mediastinum, Spine, Trachea.
AGRG-29	*Includes all AGRG-9 plus:* Aortic Arch, Carina, Cavoatrial Junction, SVC, Upper Mediastinum, Left/Right Apical Zone, Left/Right Mid Lung Zone, Left/Right Lower Lung Zone, Left/Right Upper Lung Zone, Left/Right Hilar Structures, Left/Right Clavicle, Left/Right Hemidiaphragm, Right Atrium.
AGRG-38	*Includes all AGRG-29 plus:* Left/Right Arm, Left/Right Breast, Left/Right Chest Wall, Left/Right Shoulder, Neck.

the explicit grounding tasks in CURE provide a more robust supervision signal for learning to describe radiological findings than standard text-only fine-tuning.

Benefits of Hybrid Inference. The standalone GRG prompt produces concise reports but yields lower recall (R-Mi: 0.376). Combining this global summary with fine-grained anatomical descriptions (**AGRG+GRG**) consistently yields the strongest empirical balance in our experiments. Specifically, the **AGRG-29 + GRG** configuration achieves the highest **RaTEScore** (**0.597**) in the table. Furthermore, this configuration achieves a CheXbert F1-Macro of **0.415**, marginally outperforming the state-of-the-art model CXRMate-RRG24 [33] (0.414), the winner of a recent report generation competition. This indicates that fusing a holistic global grounded report with specific, visually grounded regional descriptions is a promising strategy to bridge the gap between precision and recall in radiology report generation.

Additional Comparison with Baselines. CURE demonstrates strong performance against specialized baselines. While **CXRMate-RRG24** [33] retains the top performance on RadGraph F1 (0.255 vs. 0.176), CURE outperforms it on **RaTEScore** (0.597 vs. 0.577) and CheXbert Cosine Similarity (0.792 vs. 0.764). The strong performance on RaTEScore—a recently proposed metric designed to assess medical entities and robustness to synonyms—highlights CURE's ability to generate clinically relevant content. Furthermore, the hybrid CURE configurations consistently surpass the MAIRA-2 baseline in semantic and factual consistency metrics; specifically, **AGRG-29 + GRG** achieves a CXRFEScore of 0.655 (compared to MAIRA-2's 0.603) and a RaTEScore of 0.597 (compared to MAIRA-2's 0.496).

9.4. Extended Hallucination Analysis

To provide a holistic view of model reliability, we extend the hallucination analysis from the main paper to the full spectrum of the Chest ImaGenome schema. As detailed in the anatomical configurations (Table 17), we focus specifically on the locations for which text supervision is available (typically accompanied by bounding boxes). This selection excludes locations with bounding-box-only supervision, resulting in a comprehensive evaluation set of 38 anatomical regions. Tables 18 and 19 present the performance breakdown across these locations.

Methodology. Unlike the focused analysis in the main paper, this supplementary evaluation covers all 38 anatomical regions present in the training set with text supervision (out of 45 total locations in the schema). For each anatomy, we sampled 300 instances from the test set. To assess MAIRA-2, we utilized its **phrase grounding** capability, prompting the model with the specific anatomical name (e.g., "left clavicle") to elicit a grounded response using its official prompt template available at `https://huggingface.co/microsoft/maira-2`.

We employed `gemini-2.5-flash-lite` as an automated clinical judge to compare the anatomy-specific generation (GEN) against the full ground-truth report (GT). Using a Chain-of-Thought (CoT) prompting strategy, the judge evaluated:

1. **Correctness:** Does GEN successfully retrieve findings present in GT?
2. **Hallucination:** Does GEN invent findings not supported by GT?
3. **NLI Consistency:** What is the logical relationship (Entailment, Contradiction, Neutral) between GEN and GT?

Evaluation Prompt. The exact prompt utilized for the automated judge is provided below. It enforces an independent extraction step before comparison to minimize reasoning errors.

```
You are an expert radiologist. Your task is to compare
     a short anatomy-specific mini-report [GEN] against
     a full image ground-truth report [GT], where [GT]
     was generated by a radiologist over the entire
     image, whereas [GEN] was generated by a model over
     a specific anatomical location. You will assess
     the degree of hallucination and contradiction in
     [GEN] compared to [GT].

First, independently assess each report:
- If [GT] explicitly affirms the presence of any
     abnormality, set "gt_has_abnormalities" to "yes".
     Otherwise, set it to "no".
- If [GT] explicitly affirms the presence of any
     medical device (e.g., pacemaker, catheter, wires),
     set "gt_has_devices" to "yes". Otherwise, set it
     to "no".
- If [GEN] explicitly affirms the presence of any
     abnormality, set "gen_has_abnormalities" to "yes".
     Otherwise, set it to "no".
```

```
- If [GEN] explicitly affirms the presence of any
    medical device (e.g., pacemaker, catheter, wires),
    set "gen_has_devices" to "yes". Otherwise, set it
    to "no".

Next, perform the comparison based on [GT]:
- If [GEN] affirms the presence of an abnormality and
    this is clearly supported or reasonably suggested
    by [GT], set "gen_has_correct_abnormalities" to
    "yes". Otherwise, set it to "no".
- If [GEN] affirms the presence of an abnormality that
    is NOT affirmed nor supported by [GT], set
    "gen_has_hallucinated_abnormalities" to "yes".
    Otherwise, set it to "no".
- If [GEN] affirms the presence of a device that is
    clearly supported or reasonably suggested by [GT],
    set "gen_has_correct_devices" to "yes". Otherwise,
    set it to "no".
- If [GEN] affirms the presence of a device that is NOT
    affirmed nor supported by [GT], set
    "gen_has_hallucinated_devices" to "yes".
    Otherwise, set it to "no".
- Natural Language Inference:
   - If [GEN] makes at least one explicit statement
       that is clearly contradicted by [GT], set
       "nli_status" to "contradiction".
   - If all of [GEN]'s explicit statements are
       reasonably supported by [GT], set "nli_status"
       to "entailment".
   - Otherwise, set "nli_status" to "neutral".

You must respond ONLY with a single, valid JSON object
    in the following format. Do not add any text
    before or after the JSON object.

{
   "reason": "A detailed explanation of your reasoning
       for the comparison. Include a brief explanation
       of why you made your choices for each field.
       Focus on what is explicitly stated in [GEN] and
       [GT]. Do not make any assumptions about what is
       not explicitly stated.",
   "gt_has_abnormalities": "yes" | "no",
   "gt_has_devices": "yes" | "no",
   "gen_has_abnormalities": "yes" | "no",
   "gen_has_devices": "yes" | "no",
   "gen_has_correct_abnormalities": "yes" | "no",
   "gen_has_hallucinated_abnormalities": "yes" | "no",
   "gen_has_correct_devices": "yes" | "no",
   "gen_has_hallucinated_devices": "yes" | "no",
   "nli_status": "contradiction" | "entailment" |
       "neutral"
}
```

Results and Analysis. The comprehensive breakdown in Tables 18 and 19 highlights distinct behavioral profiles:

- **MAIRA-2 and NLI Neutrality:** MAIRA-2 exhibits a notably high *Neutral* NLI rate (73.2%) and low Abnormality Correctness (3.6%). This is the behavior one would generally expect from MAIRA-2's phrase grounding formulation: when prompted with an anatomical phrase, the model frequently outputs the phrase verbatim with coordinates, without adding descriptive adjectives. Since the output merely identifies the anatomy without making a clinical claim, the NLI judge correctly labels the relationship to the ground truth as *Neutral*. However, we observe that MAIRA-2 does occasionally append additional descriptions. When this occurs, it is highly prone to hallucination (14.9% rate), suggesting that deviations from the standard localization behavior often result in factual errors rather than useful clinical insights.
- **Sensitivity-Specificity Trade-off:** CURE demonstrates a favorable shift in the trade-off between Sensitivity (the ability to correctly identify abnormalities) and Specificity (the ability to avoid false positives). In this context, we associate *Abnormality Correctness* with Sensitivity and *Hallucination Rate* with the inverse of Specificity. CURE maintains a hallucination rate comparable to MAIRA-2 (15.2% vs. 14.9%) while achieving a massive improvement in Correctness (17.8% vs 3.6%, an approximately $5\times$ increase). This indicates that CURE's generations are more clinically useful and aligned with radiologist findings (Entailment: **43.2%** vs. 9.3%), rather than defaulting to the "safe silence" of simple object localization.
- **Anatomical Specificity:** CURE demonstrates remarkable robustness on structures where MAIRA-2 fails. For instance, on the **Left and Right Clavicles**, MAIRA-2's attempts to describe the region result in hallucination rates of 59.0% and 62.7%, respectively. This likely reflects a bias in MAIRA-2's training distribution, where mentions of the clavicles were presumably highly correlated with fractures, leading the model to hallucinate pathology even when performing a grounding task. In contrast, CURE reduces this hallucination rate to 1.0% in both cases.
- **Device Recognition:** CURE significantly outperforms MAIRA-2 in identifying medical devices (**14.0%** Correctness vs. 1.3%). It is important to note that MAIRA-2's low performance here is largely a consequence of the task formulation. MAIRA-2 is designed for specific tasks, and when performing phrase grounding using the standard prompt, it effectively localizes the structure but does not spontaneously describe the presence of devices (e.g., catheters). It is not designed as a flexible instruction-following model that can be prompted to exhaustively list findings. CURE, conversely, is trained on the AGRG objective to inherently describe the contents of the anatomical region, leading to stronger detection rates in device-heavy regions like the *Cavoatrial Junction* (45.3% Correctness).

Qualitative Example. Table 20 provides a concrete instance of the evaluation protocol applied to the *Left Clavicle*, validating the statistical trends observed above. In this scenario, the ground truth explicitly affirms that "Bony structures are intact." MAIRA-2, reflecting the high hallucination bias observed for this anatomy ($\approx 60\%$), generates a specific but incorrect finding: "Left clavicle fracture is noted." The automated judge correctly identifies this incompatibility, marking it as a *Contradiction* and a *Hallucination*. In contrast, CURE correctly generates a negative finding ("No acute osseous abnormalities"), which the

judge recognizes as supported by the ground truth (*Entailment*). This example illustrates the robustness of the automated judge in discerning clinical nuances and highlights the tangible quality improvement achieved by CURE in avoiding specific anatomical hallucinations.

10. Additional Qualitative Analysis

To provide further insight into model behavior, this section includes additional qualitative examples from both the Grounded Report Generation (GRG) and Anatomy Grounded Report Generation (AGRG) tasks.

Grounded Report Generation on VinDr-CXR. Figure 6 presents two additional test samples from the VinDr-CXR dataset. These examples highlight the difficulty of generating reports for complex cases with dense annotations. In these specific instances, both models capture salient clinical features, though challenges remain in fully recovering all localized findings present in the ground truth. Quantitative metrics for these samples remain comparable, with CURE showing a slight advantage in semantic similarity (CheXbert Cosine) and localization (IoU) in both examples.

Anatomy Grounded Report Generation on MIMIC-CXR. Figure 7 displays four examples of anatomy-specific generation. In this task, the model is prompted to locate and describe a specific region. The qualitative results highlight differences in robustness against hallucinations. For instance, in the second row (Left Clavicle), MAIRA-2 incorrectly predicts a fracture. In contrast, CURE accurately focuses its description on the placement of the endotracheal tube—which is the primary finding in the ground truth—avoiding the fracture hallucination and achieving significantly better localization (IoU 0.627 vs 0.206).

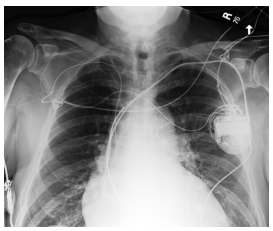
Table 18. **Per-anatomy comparison of MAIRA-2 and CURE performance on the Chest ImaGenome subset (Part 1/2).** Each anatomical region reports hallucination and correctness rates (%) for abnormalities and medical devices, along with Natural Language Inference (NLI) consistency metrics: Contradiction (Cont.) and Entailment (Ent.). Lower hallucination and contradiction values, together with higher correctness and entailment, indicate greater clinical agreement between the generated and ground-truth reports. See Table 19 for continuation.

Anatomy	Model	Abn. Halluc. ↓	Abn. Corr. ↑	Dev. Halluc. ↓	Dev. Corr. ↑	Contra. ↓	Entail. ↑	Neutral	N
Abdomen	MAIRA-2	24.0%	**6.0%**	**3.0%**	0.7%	35.7%	16.3%	48.0%	300
	CURE	**2.0%**	1.7%	14.3%	**33.0%**	**9.0%**	**54.3%**	36.7%	300
Aortic Arch	MAIRA-2	**1.3%**	0.3%	**0.0%**	0.0%	**0.3%**	16.3%	83.3%	300
	CURE	64.7%	**26.3%**	2.0%	**1.7%**	6.3%	**26.0%**	67.7%	300
Cardiac Silhouette	MAIRA-2	**2.0%**	2.7%	**0.0%**	0.3%	**8.0%**	25.0%	67.0%	300
	CURE	25.7%	**26.3%**	1.0%	**6.3%**	27.7%	**47.3%**	25.0%	300
Carina	MAIRA-2	**0.0%**	1.0%	**0.0%**	0.0%	**1.0%**	5.7%	93.3%	300
	CURE	6.0%	**2.3%**	41.0%	**14.7%**	34.0%	**11.0%**	55.0%	300
Cavoatrial Junction	MAIRA-2	**0.0%**	0.0%	**0.0%**	0.3%	**0.3%**	12.3%	87.3%	300
	CURE	6.3%	**5.7%**	32.3%	**45.3%**	15.0%	**47.0%**	38.0%	300
Left Apical Zone	MAIRA-2	7.0%	2.0%	**0.0%**	0.7%	**3.7%**	3.7%	92.7%	300
	CURE	**2.0%**	**10.3%**	**0.0%**	**4.0%**	14.0%	**61.0%**	25.0%	300
Left Arm	MAIRA-2	21.3%	2.0%	**0.7%**	0.7%	**13.3%**	6.7%	80.0%	300
	CURE	**12.7%**	**18.7%**	1.7%	**7.7%**	23.7%	**42.7%**	33.7%	300
Left Breast	MAIRA-2	27.0%	3.7%	**1.3%**	0.7%	**13.7%**	6.0%	80.3%	300
	CURE	**10.3%**	**20.0%**	**1.3%**	**7.7%**	21.0%	**44.3%**	34.7%	300
Left Chest Wall	MAIRA-2	19.3%	4.7%	11.0%	8.3%	40.0%	9.3%	50.7%	300
	CURE	**4.3%**	**18.0%**	**5.3%**	**28.0%**	**11.7%**	**68.3%**	20.0%	300
Left Clavicle	MAIRA-2	59.0%	2.3%	**0.0%**	0.3%	22.7%	5.0%	72.3%	300
	CURE	**1.0%**	**4.3%**	8.3%	**11.3%**	**7.0%**	**32.7%**	60.3%	300
Left Costophrenic Angle	MAIRA-2	**2.0%**	1.0%	**0.0%**	0.0%	**2.0%**	1.3%	96.7%	300
	CURE	9.3%	**18.7%**	**0.0%**	**2.0%**	26.3%	**53.7%**	20.0%	300
Left Hemidiaphragm	MAIRA-2	**0.7%**	0.7%	**0.0%**	0.0%	**0.3%**	4.3%	95.3%	300
	CURE	11.7%	**9.0%**	8.0%	**23.3%**	6.0%	**33.7%**	60.3%	300
Left Hilar Structures	MAIRA-2	17.7%	2.7%	**0.0%**	0.7%	**6.3%**	5.0%	88.7%	300
	CURE	**9.7%**	**31.3%**	**0.0%**	**1.3%**	28.3%	**52.3%**	19.3%	300
Left Lower Lung Zone	MAIRA-2	**11.7%**	10.0%	**0.0%**	0.0%	**8.3%**	11.7%	80.0%	300
	CURE	25.3%	**40.0%**	**0.0%**	**5.7%**	17.3%	**53.0%**	29.7%	300
Left Lung	MAIRA-2	12.0%	9.3%	0.3%	3.0%	56.3%	21.7%	22.0%	300
	CURE	**7.0%**	**12.0%**	**0.0%**	**3.3%**	**32.3%**	**41.7%**	26.0%	300
Left Mid Lung Zone	MAIRA-2	**11.0%**	4.7%	**0.0%**	0.0%	**4.7%**	4.3%	91.0%	300
	CURE	61.0%	**37.7%**	**0.0%**	**2.7%**	40.3%	**29.3%**	30.3%	300
Left Shoulder	MAIRA-2	31.3%	2.7%	**1.3%**	0.3%	**15.7%**	4.7%	79.7%	300
	CURE	**10.3%**	**20.0%**	5.7%	**8.0%**	22.7%	**44.0%**	33.3%	300
Left Upper Lung Zone	MAIRA-2	**15.3%**	8.3%	**0.0%**	0.0%	**8.0%**	7.3%	84.7%	300
	CURE	38.7%	**38.3%**	**0.0%**	**2.3%**	15.7%	**38.3%**	46.0%	300
Mediastinum	MAIRA-2	23.7%	**14.3%**	**0.0%**	4.3%	67.0%	14.0%	19.0%	300
	CURE	**5.3%**	6.0%	12.3%	**33.7%**	**20.7%**	**49.7%**	29.7%	300

Table 19. **Per-anatomy comparison of MAIRA-2 and CURE performance on the Chest ImaGenome subset (Part 2/2).** Continuation of Table 18. Each row corresponds to an anatomical region evaluated for abnormality and device hallucination (%) and correctness (%), as well as NLI-based Contradiction (Cont.) and Entailment (Ent.) rates. Lower hallucination and contradiction, and higher correctness and entailment, reflect more clinically faithful and factually consistent report generation.

Anatomy	Model	Abn. Halluc. ↓	Abn. Corr. ↑	Dev. Halluc. ↓	Dev. Corr. ↑	Contra. ↓	Entail. ↑	Neutral	N
Neck	MAIRA-2	17.0%	**3.7%**	**2.7%**	1.7%	58.3%	4.3%	37.3%	300
	CURE	**1.0%**	2.0%	27.0%	**62.0%**	**20.7%**	**41.7%**	37.7%	300
Right Apical Zone	MAIRA-2	7.3%	3.7%	**0.0%**	2.0%	**5.7%**	6.3%	88.0%	300
	CURE	**6.3%**	**13.0%**	**0.0%**	**8.0%**	20.0%	**54.0%**	26.0%	300
Right Arm	MAIRA-2	21.7%	2.0%	**1.7%**	1.0%	**16.3%**	5.3%	78.3%	300
	CURE	**7.0%**	**16.3%**	2.3%	**5.3%**	23.0%	**40.0%**	37.0%	300
Right Atrium	MAIRA-2	**0.7%**	0.3%	**0.0%**	0.3%	**1.0%**	11.3%	87.7%	300
	CURE	8.3%	**3.3%**	37.0%	**45.3%**	12.3%	**39.7%**	48.0%	300
Right Breast	MAIRA-2	21.0%	1.0%	**1.3%**	1.0%	**11.3%**	5.7%	83.0%	300
	CURE	**9.7%**	**19.3%**	3.0%	**4.7%**	22.0%	**41.3%**	36.7%	300
Right Chest Wall	MAIRA-2	23.0%	6.0%	**5.0%**	3.7%	39.7%	10.7%	49.7%	300
	CURE	**2.0%**	**23.7%**	6.0%	**20.0%**	**12.7%**	**67.0%**	20.3%	300
Right Clavicle	MAIRA-2	62.7%	0.3%	**0.0%**	0.7%	20.3%	1.7%	78.0%	300
	CURE	**1.0%**	**4.3%**	4.3%	**8.7%**	**7.3%**	**27.7%**	65.0%	300
Right Costophrenic Angle	MAIRA-2	**1.3%**	4.7%	**0.0%**	0.3%	**0.3%**	5.7%	94.0%	300
	CURE	10.7%	**25.0%**	0.3%	**3.3%**	25.3%	**49.0%**	25.7%	300
Right Hemidiaphragm	MAIRA-2	**0.7%**	0.0%	**0.0%**	0.0%	**0.0%**	6.3%	93.7%	300
	CURE	8.7%	**8.7%**	7.3%	**21.7%**	8.7%	**45.3%**	46.0%	300
Right Hilar Structures	MAIRA-2	17.0%	2.7%	**0.0%**	0.7%	**7.0%**	6.3%	86.7%	300
	CURE	**13.0%**	**32.3%**	**0.0%**	**2.3%**	31.3%	**49.0%**	19.7%	300
Right Lower Lung Zone	MAIRA-2	**8.0%**	5.3%	**0.0%**	0.0%	**6.0%**	5.3%	88.7%	300
	CURE	25.3%	**43.7%**	**0.0%**	**4.0%**	18.7%	**53.3%**	28.0%	300
Right Lung	MAIRA-2	**10.7%**	9.7%	**0.0%**	3.0%	53.7%	29.3%	17.0%	300
	CURE	11.7%	**21.3%**	0.3%	**7.3%**	**27.0%**	**46.0%**	27.0%	300
Right Mid Lung Zone	MAIRA-2	**8.7%**	2.0%	**0.0%**	0.0%	**6.0%**	2.3%	91.7%	300
	CURE	66.0%	**33.0%**	**0.0%**	**2.7%**	43.3%	**23.0%**	33.7%	300
Right Shoulder	MAIRA-2	32.0%	1.3%	**0.7%**	0.0%	**16.3%**	4.7%	79.0%	300
	CURE	**8.7%**	**22.7%**	10.0%	**16.3%**	23.7%	**45.3%**	31.0%	300
Right Upper Lung Zone	MAIRA-2	**9.7%**	3.0%	**0.0%**	0.0%	**6.3%**	3.3%	90.3%	300
	CURE	47.0%	**36.0%**	**0.0%**	**4.0%**	24.7%	**31.7%**	43.7%	300
Spine	MAIRA-2	12.7%	6.0%	5.7%	**4.7%**	38.3%	13.0%	48.7%	300
	CURE	**6.3%**	**11.7%**	**0.7%**	2.7%	**3.3%**	**41.7%**	55.0%	300
SVC	MAIRA-2	9.3%	1.0%	**4.0%**	5.0%	**11.0%**	19.0%	70.0%	300
	CURE	**5.7%**	**1.7%**	34.3%	**42.0%**	17.7%	**34.7%**	47.7%	300
Trachea	MAIRA-2	**7.7%**	0.3%	**0.0%**	0.0%	26.3%	16.0%	57.7%	300
	CURE	19.3%	**2.0%**	21.0%	**21.7%**	**24.3%**	**19.7%**	56.0%	300
Upper Mediastinum	MAIRA-2	10.0%	5.7%	**0.3%**	4.0%	33.0%	16.3%	50.7%	300
	CURE	**7.7%**	**10.7%**	4.3%	**8.7%**	**21.7%**	**61.0%**	17.3%	300
Mean	**MAIRA-2**	**14.9%**	3.6%	**1.0%**	1.3%	**17.5%**	9.3%	73.2%	300
Mean	**CURE**	15.2%	**17.8%**	7.7%	**14.0%**	20.2%	**43.2%**	36.6%	300

Table 20. **Qualitative Example of the Automated Evaluation Protocol.** We employ `gemini-2.5-flash-lite` to perform detailed hallucination and Natural Language Inference (NLI) analysis, utilizing the prompt defined in Section 9.4. This case illustrates the evaluation for the *Left Clavicle*. The automated judge compares the model-generated anatomy-specific report (GEN) against the full ground-truth report (GT). MAIRA-2 hallucinates a fracture, leading to a *Contradiction*, whereas CURE correctly identifies the lack of abnormalities, resulting in *Entailment*.



Target Anatomy: Left Clavicle

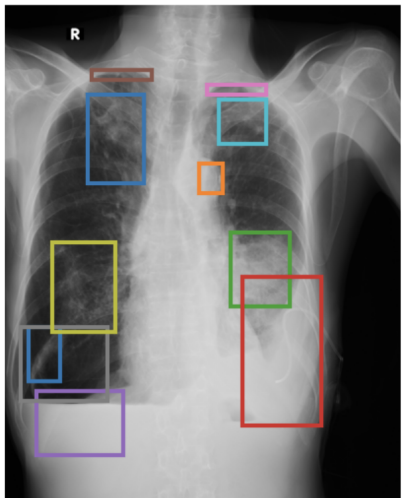
Context: Ground Truth (GT) Report

"AP upright portable chest radiograph is obtained. A left chest wall pacer device is again seen with lead tips extending into the right atrium and ventricle. Abandoned pacing leads are also seen in the right chest wall, extending into the right heart, not significantly changed. The heart is mildly enlarged. The lungs appear clear without definite signs of pneumonia or CHF. No large effusion or pneumothorax is seen. The overall cardiomediastinal silhouette is stable. ***Bony structures are intact****. No acute findings in the chest. Stable mild cardiomegaly. Multiple pacer wires are unchanged in position."*

Model	Generated Output & Gemini 2.5 Flash-Lite's Evaluation

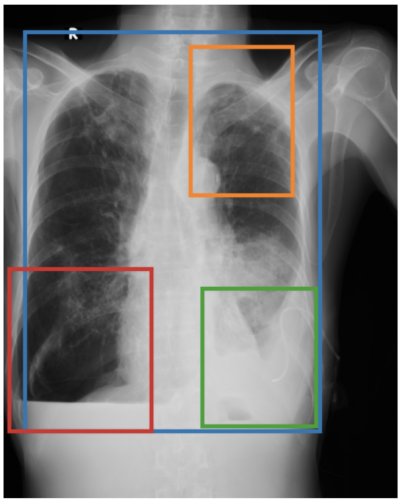
MAIRA-2	**Generation:**
	`Left clavicle fracture is noted. [0.645, 0.145, 0.995, 0.345]`
	Gemini 2.5 Flash-Lite's Evaluation:
	`{ "reason": "The Ground Truth ([GT]) report describes a left chest wall pacer device, abandoned pacing leads in the right chest wall, mild cardiomegaly, and stable findings, but it explicitly states that 'Bony structures are intact' and 'No acute findings in the chest.' The Mini-Report ([GEN]) states 'Left clavicle fracture is noted.' Since the GT explicitly states bony structures are intact and no acute findings are seen, the mention of a left clavicle fracture in [GEN] is a clear contradiction to the findings in [GT].\n \n - gt_has_abnormalities: Yes, due to mild cardiomegaly.\n - gt_has_devices: Yes, due to the presence of pacer devices and leads.\n - gen_has_abnormalities: Yes, a fracture is an abnormality.\n - gen_has_devices: No devices are mentioned in [GEN].\n - gen_has_correct_abnormalities: No, the fracture is not supported; in fact, bony structures are reported as intact.\n - gen_has_hallucinated_abnormalities: Yes, the fracture is not mentioned in [GT] and contradicts the statement that bony structures are intact.\n - gen_has_correct_devices: No, [GEN] does not mention devices.\n - gen_has_hallucinated_devices: No, [GEN] does not mention devices.\n - nli_status: Contradiction, because [GEN] reports a fracture while [GT] reports bony structures as intact and no acute findings.", "gt_has_abnormalities": "yes", "gt_has_devices": "yes", "gen_has_abnormalities": "yes", "gen_has_devices": "no", "gen_has_correct_abnormalities": "no", "gen_has_hallucinated_abnormalities": "yes", "gen_has_correct_devices": "no", "gen_has_hallucinated_devices": "no", "nli_status": "contradiction" }`

CURE	**Generation:**
	`Location of the left clavicle: [0.75,0.28,0.36,0.19]. Description: No acute osseous abnormalities.`
	Gemini 2.5 Flash-Lite's Evaluation:
	`{ "reason": "The ground truth report [GT] states, \"Bony structures are intact\" and \"No acute findings in the chest,\" which implies no acute osseous abnormalities. The mini-report [GEN] states, \"No acute osseous abnormalities.\" This statement is directly supported by [GT].\n \nFor the initial assessment:\n - [GT] mentions a pacer device and leads, so gt_has_devices is 'yes'. It mentions mild cardiomegaly, so gt_has_abnormalities is 'yes'.\n - [GEN] mentions no abnormalities, so gen_has_abnormalities is 'no'. [GEN] mentions no devices, so gen_has_devices is 'no'.\n \nComparison :\n - [GEN] does not affirm any abnormality, so gen_has_correct_abnormalities and gen_has_hallucinated_abnormalities are 'no'.\n - [GEN] does not affirm any device, so gen_has_correct_devices and gen_has_hallucinated_devices are 'no'.\n - Since the statement in [GEN] (\"No acute osseous abnormalities\") is directly supported by the statement in [GT] (\"Bony structures are intact\"), the NLI status is 'entailment'.", "gt_has_abnormalities": "yes", "gt_has_devices": "yes", "gen_has_abnormalities": "no", "gen_has_devices": "no", "gen_has_correct_abnormalities": "no", "gen_has_hallucinated_abnormalities": "no", "gen_has_correct_devices": "no", "gen_has_hallucinated_devices": "no", "nli_status": "entailment" }`



(a) Ground Truth

Atelectasis [0.06,0.65,0.14,0.76]. Calcification [0.49,0.32,0.55,0.38]. Consolidation [0.57,0.46,0.72,0.61]. Pleural effusion [0.60,0.55,0.80,0.85] [0.08,0.78,0.30,0.91]. Pleural thickening [0.22,0.13,0.37,0.15] [0.51,0.16,0.66,0.18]. Pneumothorax [0.04,0.65,0.26,0.80]. Pulmonary fibrosis [0.12,0.48,0.28,0.66] [0.54,0.19,0.66,0.28] [0.21,0.18,0.35,0.36]. Tuberculosis. Other diseases.

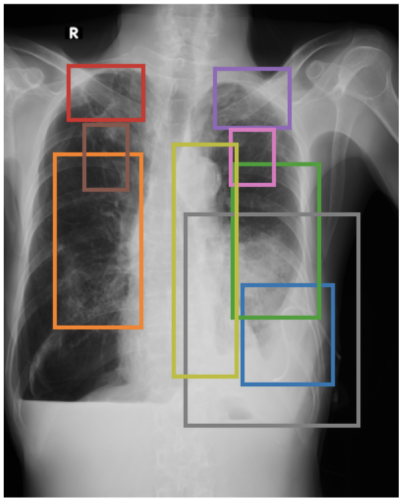


(b) MAIRA-2

The heart size is normal. Lung fields are hyperaerated bilaterally. [0.055, 0.055, 0.805, 0.865] There is pleural calcification in the left upper lobe. [0.475, 0.085, 0.735, 0.385] There is blunting of the left cp angle. [0.505, 0.575, 0.795, 0.855] There is pleural calcification in the right lower hemithorax. [0.015, 0.535, 0.375, 0.865] There is blunting of the right cp angle.

Metrics:

IoU:	0.311
CheXbert Cos:	0.767
CheXbert Acc:	0.571
CXRFEScore:	0.416

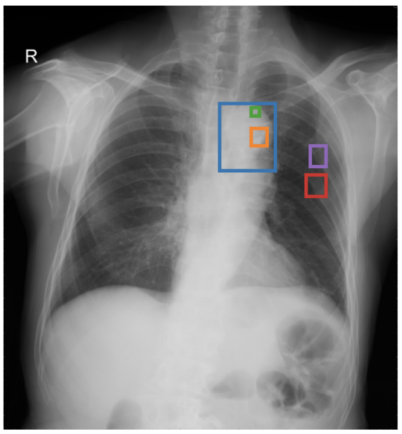


(c) CURE

Left pleural effusion without significant changes [0.72,0.67,0.23,0.20]. Chronic changes in the lung parenchyma [0.24,0.48,0.22,0.35] [0.69,0.48,0.22,0.31]. Biapical pleuroparenchymal thickening [0.26,0.18,0.19,0.11] [0.63,0.19,0.19,0.12]. Calcified granulomas [0.26,0.31,0.11,0.13] [0.63,0.31,0.11,0.11]. Changes from left mastectomy [0.68,0.64,0.44,0.43]. Aortic elongation [0.51,0.52,0.16,0.47]

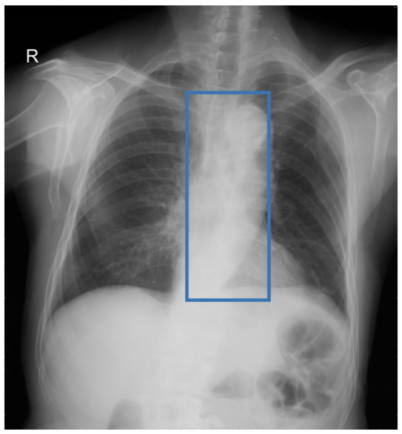
Metrics:

IoU:	**0.320**
CheXbert Cos:	**0.834**
CheXbert Acc:	**0.643**
CXRFEScore:	**0.523**



(d) Ground Truth

Aortic enlargement [0.55,0.23,0.69,0.39]. Calcification [0.63,0.29,0.67,0.33] [0.63,0.24,0.65,0.26]. Rib fracture [0.77,0.40,0.82,0.45] [0.78,0.33,0.82,0.38]. Other diseases.

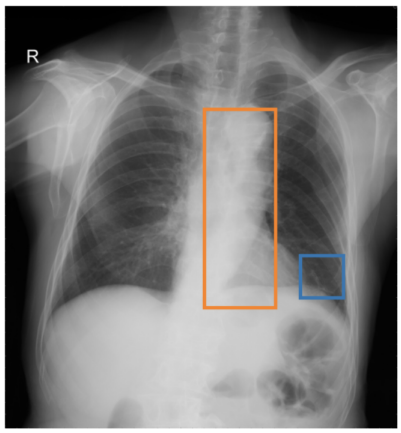


(e) MAIRA-2

Normal cardiac silhouette. The lung fields are clear. No evidence of hilar adenopathy. No evidence of mediastinal adenopathy. Skeletal structures are unremarkable. The aorta is tortuous. [0.465, 0.205, 0.675, 0.695]

Metrics:

IoU:	0.182
CheXbert Cos:	0.738
CheXbert Acc:	**0.786**
CXRFEScore:	0.316



(f) CURE

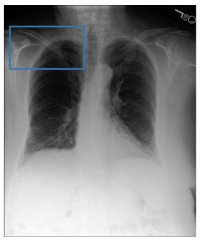
Left basal laminar atelectasis [0.81,0.64,0.11,0.10]. Aortic elongation [0.60,0.48,0.18,0.47]. No other relevant findings

Metrics:

IoU:	**0.199**
CheXbert Cos:	**0.755**
CheXbert Acc:	**0.786**
CXRFEScore:	**0.350**

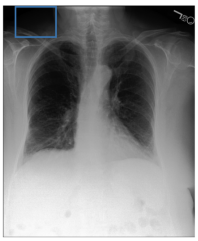
Figure 6. Qualitative comparison of two examples of grounded report generation (GRG) from the **VinDr-CXR** dataset. The **top row (a-c)** shows the first example, and the **bottom row (d-f)** shows the second example. For each, the left column represents ground-truth annotations, the middle is MAIRA-2, and the right is CURE. The colored coordinates in the text correspond to bounding boxes drawn in the images. Per-sample metrics are provided below the predicted reports.

(a) Ground Truth



Loc: Right Clavicle
Osseous structures are diffusely demineralized. [0.03, 0.09, 0.42, 0.28]

(b) MAIRA-2

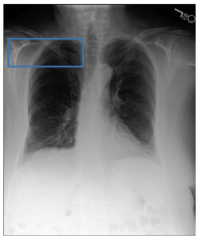


Right clavicle is fractured. [0.07, 0.01, 0.28, 0.14]

Metrics:

IoU:	0.097
CheXbert Cos:	0.459
CheXbert Acc:	0.857
CXRFEScore:	0.119

(c) CURE

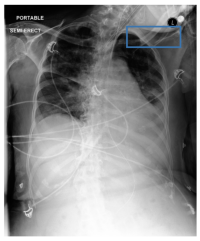


Location of the right clavicle: [0.22,0.21,0.39,0.12].
Description: No acute osseous abnormalities.

Metrics:

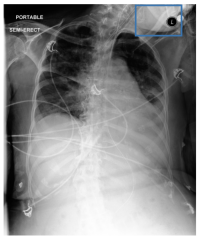
IoU:	**0.643**
CheXbert Cos:	**0.882**
CheXbert Acc:	**1.000**
CXRFEScore:	**0.513**

(d) Ground Truth



Loc: Left Clavicle
ETT tip at clavicle margin. No bone abnormalities. [0.64, 0.09, 0.92, 0.18]

(e) MAIRA-2

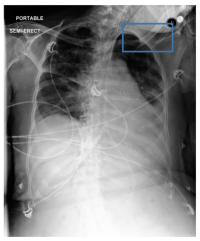


Left clavicle fracture. [0.69, 0.01, 0.93, 0.14]

Metrics:

IoU:	0.206
CheXbert Cos:	0.303
CheXbert Acc:	0.786
CXRFEScore:	0.404

(f) CURE



Location of the left clavicle: [0.75,0.14,0.26,0.12].
Description: The patient is intubated, the ETT terminating in the trachea 4 cm above the level of the carina.

Metrics:

IoU:	**0.627**
CheXbert Cos:	**0.880**
CheXbert Acc:	**1.000**
CXRFEScore:	**0.467**

(g) Ground Truth

Loc: Right Lung
There is a diffuse mild interstitial abnormality in the right lung, unchanged from prior. There is no evidence of consolidation or edema. There is no pleural effusion or pneumothorax. There is evidence of stable pulmonary hypertension and vascular engorgement. [0.086, 0.219, 0.466, 0.705]

(h) MAIRA-2

Right lung is clear.

Metrics:

IoU:	0.000
CheXbert Cos:	**0.716**
CheXbert Acc:	**0.929**
CXRFEScore:	0.530

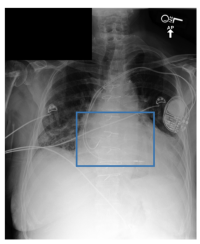
(i) CURE

Location of the right lung: [0.28,0.48,0.36,0.48].
Description: There is mild pulmonary vascular congestion. No pleural effusion or pneumothorax is seen. Mild pulmonary vascular congestion.

Metrics:

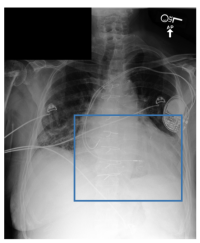
IoU:	**0.879**
CheXbert Cos:	0.699
CheXbert Acc:	**0.929**
CXRFEScore:	**0.805**

(j) Ground Truth



Loc: Cardiac Silhouette
Moderate cardiomegaly is unchanged. A left-sided dual-chamber pacemaker is present with leads in the right atrium and right ventricle. [0.375, 0.451, 0.778, 0.683]

(k) MAIRA-2

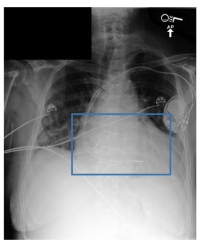


cardiac silhouette. [0.37, 0.47, 0.93, 0.84]

Metrics:

IoU:	0.411
CheXbert Cos:	0.723
CheXbert Acc:	**0.929**
CXRFEScore:	0.515

(l) CURE



Location of the cardiac silhouette: [0.61,0.59,0.51,0.26].
Description: FINDINGS: In comparison with the study of ___, there is continued enlargement of the cardiac silhouette with a dual-channel pacer device in place.

Metrics:

IoU:	**0.657**
CheXbert Cos:	**0.725**
CheXbert Acc:	**0.929**
CXRFEScore:	**0.659**

Figure 7. Qualitative comparison of Anatomy Grounded Report Generation (AGRG) on the **MIMIC-CXR** test set. We display images and reports side-by-side for compactness. Rows correspond to: **(a-c)** Right Clavicle, **(d-f)** Left Clavicle, **(g-i)** Right Lung, and **(j-l)** Cardiac Silhouette. The left block is ground-truth, middle is MAIRA-2, and right is CURE. Per-sample metrics are provided below each anatomy-grounded report.