This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Pablo Messina[1,2,3], Andrés Villa[4], Juan León Alcázar[4], Karen Sanchez[4],
Carlos Hinojosa[4], Denis Parra[1,2,3], Alvaro Soto[1,2], Bernard Ghanem[4]
[1]Pontificia Universidad Católica de Chile, [2]CENIA, [3]iHEALTH, [4]KAUST
pamessina@uc.cl, {andres.villa,juancarlo.alcazar,karen.sanchez,carlos.hinojosa}@kaust.edu.sa
{dparra,asoto}@ing.puc.cl, bernard.ghanem@kaust.edu.sa

Abstract

Medical vision–language models can automate the generation of radiology reports but struggle with accurate visual grounding and factual consistency. Existing models often misalign textual findings with visual evidence, leading to unreliable or weakly grounded predictions. We present "CURE", an error-aware curriculum learning framework that improves grounding and report quality without any additional data. CURE tunes a multimodal instructional model on phrase grounding, grounded report generation, and anatomy-grounded report generation using public datasets. The method dynamically adjusts sampling based on model performance emphasizing harder samples to improve spatial and textual alignment. CURE improves grounding accuracy by +0.35 IoU, boosts report quality by +0.192 CXRFEScore, and reduces hallucinations by 18.6%. CURE is a data-efficient framework that enhances both grounding accuracy and report reliability. Code is available at https://github.com/PabloMessina/CURE and model weights at https://huggingface.co/pamessina/medgemma-4b-it-cure.

1. Introduction

In the medical domain, Vision-language Models (VLMs) enable the automatic generation of diagnostic reports from imaging exams, thereby reducing the workload of specialists and promoting standardized diagnostic pipelines [3, 23, 38]. Despite their remarkable success in bridging visual and textual modalities [1, 10, 25, 35, 46], VLMs still face fundamental challenges when applied to specialized domains such as medicine. The factuality and reliability of VLMs remain a significant concern, as current architectures often generate outputs that are inconsistent with their visual input, a phenomenon known as hallucinations [24, 45, 48].

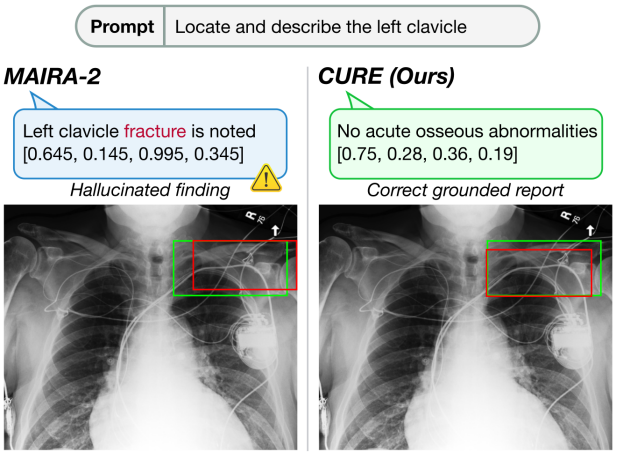


Figure 1. **False Positive Detection of Pathologies.** Given the same chest X-ray input from the MIMIC-CXR test set, both models approximate the location of the left clavicle. However, the baseline model (MAIRA-2) hallucinates a fracture (there is no fracture in the image), whereas our proposed model (CURE) generates a clinically correct and visually grounded description.

Medical VLMs [23, 38, 55] rely on large-scale, domain-specific datasets to adapt general-purpose multi-modal models for medical applications. While this standard fine-tuning strategy improves performance on medical benchmarks, the resulting models often lack visual grounding, thus limiting interpretability and increasing the risk of hallucinations [47]. For clinical adoption, it is crucial that VLMs produce factually accurate outputs and correctly ground medical findings in the relevant image regions. Strengthening factuality and grounding is therefore essential to ensure transparency, trustworthiness, and the reliable deployment of medical VLMs in real-world workflows.

To address these drawbacks, state-of-the-art models such as MAIRA-2 [3] explicitly incorporate grounded report generation to better align local visual evidence with the textual findings. For example, MAIRA-2 is trained on a diverse

set of localization-based tasks, including phrase grounding and grounded report generation. However, as shown in Figure 1, current models often exhibit a bias in which the target visual regions become overly associated with abnormal findings, leading to false positives in medical reports.

To address these limitations, we propose **CURE** (**CU**rriculum-guided Multi-task Training for **RE**liable Anatomy Grounded Report Generation), a curriculum-based learning framework that enhances the reliability and visual grounding of medical VLMs without requiring additional data. CURE improves the standard training pipeline, and obtains improved results using only a subset of publicly available chest X-ray datasets. In particular, CURE replaces the conventional finding-generation objective used in state-of-the-art methods with an anatomy-grounded report generation (AGRG) task, thus leveraging the most detailed regional annotations provided by some datasets.

Our contributions are threefold: **i)** We introduce CURE, a novel error-aware curriculum framework that dynamically adjusts sampling distributions based on model performance, enhancing visual grounding without requiring any additional training data. **ii)** Our training strategy enables the effective transfer of grounding capabilities to medical VLMs that originally lack visual grounding mechanisms. These models surpass the current state of the art in visual grounding, despite being trained on less training data. **iii)** On the Chest ImaGenome dataset, CURE achieves an 18.6% reduction in hallucinations across six key anatomical regions, substantially improving the reliability and trustworthiness of medical VLM outputs and setting a new state-of-the-art in visually grounded report generation.

2. Related Work

Medical Vision-language Models. The adoption of Vision-language models (VLMs) in the medical domain has advanced rapidly in recent years [15, 18, 31, 42]. Early approaches primarily targeted clinical report generation and captioning tasks, while more recent models such as MedPaLM [42, 43], MAIRA-2 [2], and the recently released MedGemma [39] exhibit strong multi-modal reasoning capabilities across diverse medical modalities. Despite their success, the challenge of efficiently fine-tuning these large models on heterogeneous and task-specific medical datasets remains largely unsolved. Our work builds upon this line of research by introducing a structured, error-aware training curriculum that enhances performance without requiring additional data or model parameters.

Multi-Task Learning in Medical Imaging. Multi-task learning (MTL) is a widely adopted paradigm for jointly training models on related objectives such as classification, segmentation, and report generation [8, 26, 56]. In medical imaging, MTL has been successfully applied across modalities, improving both efficiency and generalization [6, 9, 20, 30, 36, 50, 51]. For instance, Sainz *et al*. [36] leveraged MTL for breast cancer screening by jointly learning classification and detection of abnormal findings in mammography; Chen *et al*. [9] improved atrial segmentation and classification using MRI; and Weninger *et al*. [50] enhanced brain tumor segmentation by coupling detection and image reconstruction tasks in brain MRI. Despite these advances, MTL performance strongly depends on how task contributions are balanced during training [19]. Improper weighting can cause task dominance and negative transfer [30, 40, 54], degrading overall performance. To address this limitation, we propose an adaptive curriculum that dynamically schedules data exposure based on model performance, allowing the network to automatically prioritize under-performing tasks rather than relying on manually tuned loss weights.

Curriculum Learning. Curriculum learning (CL)[4, 49] trains models by gradually increasing task difficulty, starting from easier samples and progressing to harder ones, in analogy to human learning. CL has been successfully applied to medical imaging [22, 41]. For instance, Shibu *et al*. [41] introduced a MedSAM-guided CL strategy for white matter tract segmentation, where the curriculum is defined by anatomical block complexity to progressively refine spatial representations. Similarly, Li *et al*. [22] introduced a dynamic CL framework for medical image classification that derives sample difficulty from in-domain uncertainty estimates (via a Dirichlet classifier) and adapts the sampling schedule accordingly. Most CL implementations rely on predefined or heuristic difficulty measures. Instead, more recent extensions, termed self-paced learning (SPL) [21], estimate sample difficulty using the training loss at each iteration. Yet, SPL tends to repeatedly select easy samples, since data points with lower losses are consistently prioritized. To overcome this limitation, Jiang *et al*. [16] proposed self-paced curriculum learning (SPCL), which integrates prior knowledge from predefined curricula with the adaptive nature of SPL, enabling the model to leverage both task structure and feedback during learning. Building on this principle, our method employs an error-aware sampling strategy, where sample difficulty is inferred from the model's current error distribution across tasks and data subsets. Unlike prior work, our curriculum integrates both spatial and textual feedback to guide training in grounded medical vision-language models.

3. Methodology

Our curriculum-guided multi-task training framework, CURE, enhances the capabilities of medical VLMs by reformulating the training methodology without requiring additional data. CURE restructures existing datasets into a unified, fine-grained instructional format. In addition, it introduces an error-aware curriculum that dynamically adjusts sampling based on the model's performance across

Table 1. **Dataset composition and statistics.** Number of instances for each task across the training, validation, and test splits. The MIMIC-CXR dataset serves as a superset, providing the Chest ImaGenome (CIG) and MS-CXR subsets for training and its official test split for report-generation evaluation. Evaluation-only datasets are used to assess generalization performance. **AGRG** refers to Anatomy Grounded Report Generation, **PG** to Phrase Grounding, and **GRG** to Grounded Report Generation.

Dataset	Task	# Train	# Val	# Test
Training & In-Domain Evaluation				
MIMIC-CXR (CIG)	AGRG *Locate*	8.5M‡	69.9K†	1,000*
	AGRG *Describe*	2.3M‡	18.8K†	
	AGRG *Loc., Desc.*	2.1M‡	17.4K†	
MIMIC-CXR (MS-CXR)	PG	815‡	169†	176
MIMIC-CXR (Test Split)	Report Generation	—	—	3,155
PadChest-GR	GRG	3,185‡	455†	915
	PG	8,871**‡	1,297†**	1,238
Zero-Shot Generalization				
VinDr-CXR	GRG	—	—	3,000
	PG	—	—	2,108

*A sampled subset of 1,000 from 123,319 total available instances.
**Includes both original report phrases and fine-grained labels as phrases.
†Each curriculum cycle evaluates a stratified random subset of the validation set.
‡Due to computational constraints, only a fraction of the total available instances are used. With 9,000 training steps and an effective batch size of 25, a maximum of 225,000 instances (≈1.74% of the total) are processed.

datasets and anatomical categories. This reformulated training pipeline enables the model to focus on challenging samples and underperforming regions progressively.

3.1. Error-Aware Curriculum Learning

Our training setup involves a collection of k distinct data sources, $\mathcal{D} = \{D_1, D_2, \ldots, D_k\}$. Each data source ($D_i$) corresponds to a specific dataset and its original supervision task, such as phrase grounding (PG) in MS-CXR or anatomy grounded report generation (AGRG) in Chest ImaGenome. Under standard multi-task training, samples are drawn from $\mathcal{D}$ in proportion to the dataset size ($|D_i|$), the sampling probability for D_i is defined as: $p_i = |D_i|/|\mathcal{D}|$.

However, medical imaging datasets are inherently imbalanced in terms of relative dataset sizes and in the distribution of anatomical or semantic classes within each dataset. The second source of imbalance is further emphasized due to variations in patient demographics, acquisition protocols, and the natural frequency of clinical conditions [37].

As shown in Table 1, our data composition reflects this imbalance: over 12.9M instances from Chest ImaGenome dominate the much smaller MS-CXR (815 PG instances) and PadChest-GR (∼12k instances) datasets. Beyond the imbalance between datasets, each dataset also exhibits substantial intra-dataset class imbalance. For example, anatomical regions in AGRG and semantic categories in PG are unevenly represented, leading the model to overfit frequent regions, neglect rare but clinically important ones, and hallucinate findings (Figure 1). Our curriculum framework addresses both sources of imbalance.

To mitigate these issues, we introduce an error-aware curriculum learning strategy [13] that dynamically adjusts sampling probabilities at two levels of granularity: (i) at the dataset level, where we re-weight the relative number of training samples per dataset, and (ii) at the class level, where we re-weight samples per anatomical region or semantic class. The curriculum proceeds over n iterative stages. Each stage consists of three steps: training, evaluation, and sample re-weighting. The sampling probability for each data source (p_i) and class is modulated by the error rate (e_i) at the current stage. The sampling probabilities estimated at stage n are used to initialize stage $n+1$, while the first stage starts with a uniform distribution *i.e.*, $p_i = p_j$ for all i, j. The overall procedure is illustrated in Figure 2.

Data Source Evaluation. To maintain computational efficiency, each stage evaluates a random, fixed-size subset from each validation source (e.g., ∼200 samples for Chest ImaGenome, ∼150 for PadChest-GR, and ∼100 for MS-CXR). For each subset, we assess boxes localization quality using Intersection over Union (IoU) and textual fidelity using CXRFEScore [29], which captures clinical semantic similarity. These metrics determine the updated sampling weights for the next stage.

Inter-Dataset Curriculum. After each evaluation stage, we compute an aggregate performance score (s_i) for each data source D_i as a weighted sum of the two metrics:

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

where α controls the trade-off between localization accuracy and semantic quality. The error for each source is defined as $e_i = 1 - s_i$. The sampling probability for the next training stage is then obtained by normalizing these errors:

$$p_i = \frac{e_i}{\sum_{j=1}^{K} e_j}. \tag{2}$$

Intra-Dataset Curriculum. Within each dataset, the curriculum operates at a finer granularity by re-weighting categories, anatomical regions, or semantic groups based on their per-class error. For phrase grounding (PG) in MS-CXR [5], we group samples by the eight original phrase classes (e.g., pneumonia, consolidation). In PadChest-GR [11], where annotations span 155 labels organized into 26 higher-level label groups (e.g., atelectasis, cardiomegaly), we apply the curriculum at the group level.

The most fine-grained application of the curriculum occurs in Chest ImaGenome [53] for the anatomy grounded report generation (AGRG) task. AGRG comprises three subtasks: *Locate*, *Describe*, and *Locate and describe*, sampled uniformly to preserve task balance. The curriculum

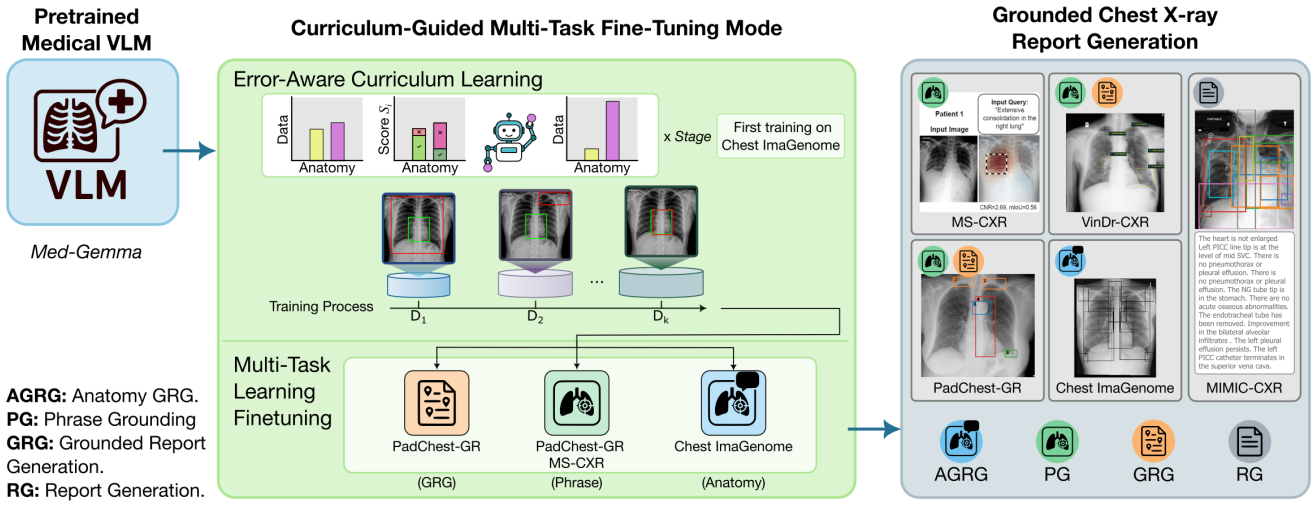


Figure 2. **Overview of CURE, our Curriculum-guided Multi-task Training Framework.** During training, the model is periodically evaluated every N steps on validation subsets from each task. Performance metrics (IoU, CXRFEScore) are calculated to identify task-level and category-level errors, which are then used to update the sampling weights in the training sampler. The cycle then resumes, allowing the model to focus more heavily on the data it finds most challenging. Evaluation of the RG task uses the official MIMIC-CXR test set, while VinDr-CXR is assessed in a zero-shot setting.

acts within each subtask by re-weighting anatomical locations according to their error.

Because annotation coverage differs across subtasks, each one has its own set of available locations (i.e., 36 with bounding box for *Locate*, 38 with text for *Describe*, and 29 with both annotations for *Locate and describe*). This formulation allows the model to focus on anatomical regions that are spatially or semantically challenging, improving both localization and descriptive quality.

For the grounded report generation (GRG) task in PadChest-GR, we do not apply intra-dataset re-weighting. These reports contain multiple co-occurring findings spanning several anatomical and pathological categories. Since there is no straightforward or well-defined way to categorize such multi-label instances for fine-grained rebalancing, we adopt uniform sampling for this task.

3.2. Fine-Grained Task Formulation

CURE re-formulates diverse tasks into a unified, fine-grained instructional format. Each training instance from the source dataset D_i is represented as a triplet `(image, instruction, response)`, allowing heterogeneous supervision (bounding boxes, phrases, anatomical labels, and descriptive sentences) to be learned under a consistent multi-task framework.

Phrase Grounding (PG). For MS-CXR and PadChest-GR, the original annotations link phrases within a report to one or more corresponding bounding boxes on the image. We transform these annotations into direct grounding instructions. For each annotated phrase, we create a sample using the template prompt `Ground the phrase: {phrase}`. When a phrase is associated with multiple bounding boxes, all of them are included in the response. The expected output is the phrase followed by one or more normalized bounding box coordinates: `phrase: [cx1, cy1, w1, h1] ... [cxn, cyn, wn, hn]`.

PadChest-GR provides $8,489$ sentences and 155 unique, fine-grained clinical labels. To enrich our training data, we exploit this dual annotation by generating additional instances: for each `(sentence, box list)` pair, we create `(label, box list)` pairs. For example, from "Minimal biapical pleural thickening" with label "apical pleural thickening" and bounding box `[cx, cy, w, h]`, we produce one instance for the sentence and another for the label. This augmentation, applied only to training and validation splits, nearly doubles the PadChest-GR PG data and helps the model ground natural descriptions and canonical clinical terms. The test set remains un-augmented.

Grounded Report Generation (GRG). In the PadChest-GR's GRG task, the model must produce a complete report in which grounded findings are explicitly linked to bounding boxes. We use the instruction: `Generate a grounded report`. The target response is the ground-truth report, where phrases corresponding to available bounding box annotations are augmented with their coordinates directly in the text. For example, an output may take the form: `{phrase1} [cx1, cy1, w1, h1]. {phrase2} [cx2, cy2, w2, h2]`, where each phrase is linked to its corresponding list of bounding boxes.

Anatomy-Grounded Report Generation (AGRG). Chest ImaGenome provides detailed scene-graphs linking anatomical locations in MIMIC-CXR frontal-view images

Table 2. **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. CURE consistently improves localization performance across all metrics and datasets, including VinDr-CXR, which was not seen during training.

Model	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
CURE	**0.552** ± 0.015	**0.495** ± 0.015	**0.453** ± 0.006	**0.438** ± 0.007	**0.243** ± 0.005	**0.205** ± 0.012

to both descriptive sentences and bounding boxes. This detailed structure is ideal for our AGRG fine-grained task decomposition as it allows for a significant expansion of the training data: a single image can yield multiple training instances, one for each annotated location. Concretely, the ∼237k images in the training split generate an average of 9 to 36 instructional instances per image, depending on the subtask, culminating in millions of viable training samples reported in Table 1. From this large pool, we create three distinct subtasks:

- **Locate:** To train pure spatial localization, we use the prompt `Locate the {location}`. The expected output follows the format: `Location of the {location}: [cx,cy,w,h]`.
- **Describe:** To train contextual description independent of localization, we use the prompt `Describe the {location}`. The target output is formatted as: `Description of the {location}: {description}`.
- **Locate and Describe:** To train the model to perform both tasks simultaneously, we prompt: `Locate and describe the {location}`. The response combines both formats: `Location of the {location}: [cx,cy,w,h]. Description: {description}`.

This multi-prompt formulation explicitly disentangles and teaches the model the diverse tasks of localization and description. This process generates over 12.9M potential training instances. However, it is important to note that due to computational constraints, only a fraction of this pool is sampled during training, as detailed in the footnote of Table 1.

4. Experiments

We evaluate CURE on Anatomy-Grounded Report Generation (AGRG), Grounded Report Generation (GRG), Phrase Grounding (PG), and traditional report generation (RG). For RG, we repurpose the AGRG and GRG setups (see Supplementary Table 9). We benchmark against state-of-the-art methods and ablate core components to assess their impact on grounding accuracy and report quality.

Datasets. CURE adopts a multi-task fine-tuning strategy on the same three publicly available chest X-ray datasets used by our baseline (MAIRA-2), enabling a fair and comparable evaluation. Chest ImaGenome [53] is a subset derived from the MIMIC-CXR [17] dataset that focuses exclusively on frontal-view X-rays. It provides scene graphs linking anatomical regions to bounding boxes and their corresponding original MIMIC-CXR report sentences, and is used for the Anatomy-Grounded Report Generation task. PadChest-GR [11] contains radiology reports with phrases explicitly grounded in bounding boxes, supporting both Grounded Report Generation and Phrase Grounding tasks. MS-CXR [5] is a smaller dataset derived from MIMIC-CXR, providing bounding box annotations for report phrases, and is used exclusively for Phrase Grounding. For zero-shot evaluation, we also include VinDr-CXR [32], using its official test split to assess the model's generalization to unseen data distributions.

Evaluation Metrics. For the *Grounded Report Generation* task, localization performance is measured using the mean Intersection over Union (IoU), while textual report quality is evaluated using CheXbert [44] metrics (such as Precision, Recall, F1-score, and Cosine Similarity), CXR-FEScore [28], RaTEScore [57], and RadGraph F1 [12]. For the *Phrase Grounding task*, we report both micro- and macro-averaged IoU to evaluate across and within categories. All reported metrics include standard deviations derived from 1000 bootstrap samples. Please refer to Supplementary Section 7 for a detailed overview of the tasks, corresponding prompt formats, and expected output structures for each dataset used in our experiments.

We use MedGemma-4B-IT [38] as the base model, fine-tuned with LoRA [14] at rank 16 in 4-bit precision. All training stages use the AdamW optimizer [27] with a consistent set of hyperparameters: an effective batch size of 25 (per-device batch size 5 with 5 gradient accumulation steps), a learning rate of 2×10^{-4}, a linear scheduler with a 0.03 warmup ratio, and gradient clipping with maximum gradient norm set to 0.3. Optimizer and scheduler states reset between phases; the initial 3000-step pre-training and subsequent 6000-step multi-task phase use separate optimizer instances, preserving only model weights. Data augmentation includes spatial transformations and contrast-limited adaptive histogram equalization (CLAHE) [58].

Baseline Methods. We compare CURE against two strong baselines. First, we evaluate MedGemma-4B-IT [38] to establish the pretrained model's baseline performance. Second, we include MAIRA-2 [2], a state-of-the-art open-source medical vision–language model that jointly learns grounding and report generation tasks.

4.1. Comparison against State-of-the-Art

Visual Grounding. Table 2 shows that CURE consistently outperforms MAIRA-2 across all metrics (Micro and Macro IoU) and datasets, including MS-CXR and PadChest-GR. CURE achieves larger relative gains on PadChest-GR than on MS-CXR, highlighting the effectiveness of the proposed error-aware reweighting strategy.

Anatomy-Grounded Report Generation. On Chest ImaGenome, CURE gains +0.35 IoU over MAIRA-2, doubling localization performance (Table 3). Since MedGemma-4B-IT lacks innate visual grounding, this confirms our pipeline successfully instills it. Furthermore, CURE surpasses MAIRA-2's IoU on GRG (Table 4), despite MAIRA-2's pretraining on the proprietary USMix dataset [3] (193K text-only, 69K grounded reports).

Report Generation Performance. We evaluate CURE across three report generation tasks: on the MIMIC-CXR test set for standard report generation, and on Chest-ImaGenome and PadChest-GR for the AGRG and GRG tasks, respectively. Table 3 presents AGRG results on a 1000 sample subset of the Chest ImaGenome test set (see Table 1). While our model's most significant improvement is in spatial grounding, it also shows enhanced performance in most text-based metrics for this fine-grained, in-domain task. CURE achieves the highest scores on F1-Mi (0.529), cosine similarity (0.691), and CXRFEScore (0.549), outperforming the baseline models.

Table 4 reports GRG performance on PadChest-GR. This task naturally favors MAIRA-2, owing to its additional training on the proprietary USMix dataset, which allows it to surpass CURE on most text-based metrics. In contrast, CURE consistently delivers higher grounding accuracy, achieving the best mean IoU (0.265). Although CURE does not outperform MAIRA-2 on text-based metrics, it substantially boosts MedGemma-4B-IT's performance in these metrics. This highlights the effectiveness of our training paradigm, particularly given that PadChest-GR (GRG) contains far fewer training instances than its version for PG and Chest ImaGenome for AGRG.

Table 5 shows performance on the MIMIC-CXR test set, comprising 3,155 frontal views. For fair comparison, all models generate reports from the same frontal view. Our CURE (AGRG) variant, which generates reports by concatenating the descriptions for 29 distinct anatomical locations, outperforms all the baselines in recall scores (R-Ma: 0.539, R-Mi: 0.749), and the combined CURE (AGRG+GRG) further improves this (R-Ma: 0.582, R-Mi: 0.781), demonstrating that our fine-grained approach effectively captures a wide range of findings. The top-performing model on most metrics is CXRMate-RRG24 [33], the winner of a recent radiology report generation competition. Its leading RadGraph F1 score is consistent with its training, which was optimized via reinforcement learning using RadGraph F1 as a reward. Comparatively, our model remains competitive on semantic similarity metrics, achieving the highest CheXbert Cosine Similarity (0.792), RaTEScore (0.597), and a close second on CXRFEScore (0.655). CURE also achieves the highest CheXbert F1 macro (0.415).

Table 3. **Results for Anatomy-Grounded Report Generation (AGRG) on Chest ImaGenome (CIG).** We report mean IoU (↑), CheXbert F1 (micro/macro) (↑), CheXbert cosine similarity (↑), and CXRFEScore (↑). **Bold** values indicate the best performance for each metric.

Model	IoU ↑	F1-Mi ↑	F1-Ma ↑	Cos. ↑	CXRFEScore ↑
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
CURE	**0.601** ± 0.008	**0.529** ± 0.017	**0.234** ± 0.018	**0.691** ± 0.009	**0.549** ± 0.011

Evaluating Hallucination and Reliability. To complement our empirical evaluation, we conduct a targeted analysis of hallucination and report consistency on the AGRG task. The evaluation is performed on a subset of the Chest ImaGenome test set, comprising 300 examples for six key anatomical locations (1800 images in total). We prompted CURE using its `Locate and describe` instruction and prompted MAIRA-2 to ground and describe the same location. We then employed Gemini 2.5 Flash Lite to perform a Natural Language Inference (NLI) comparison between each model's generated description for a specific anatomy and the full ground-truth MIMIC-CXR report.

As detailed in Table 6, CURE substantially improves reliability. On average, it reduces abnormal finding hallucinations from MAIRA-2's 26.50% to 8.78%, halves contradictions (17.44% vs. 33.22%), and more than doubles entailment (39.50% vs. 15.94%). The improvement is particularly stark for bone structures like the clavicles, where CURE's hallucination rate is only 1.00% compared to MAIRA-2's rates of over 59%.

While MAIRA-2 exhibits a slightly lower hallucination rate for a few anatomies (e.g., 'Cardiac Silhouette'), CURE consistently achieves a significantly lower contradiction rate and higher entailment rate across almost all categories. This improved performance likely stems from a key difference in the training data composition. Standard phrase grounding, used by MAIRA-2, is inherently biased towards abnormal findings. In contrast, our AGRG formulation exposes the model to both normal and abnormal descriptions for each anatomical region, leading to a more balanced and reliable generative process that mitigates the tendency to hallucinate abnormalities.

Zero-Shot Performance. To assess the generalization capabilities of CURE beyond the datasets seen during training, we further evaluate the model on the VinDr-CXR test split for both PG and GRG (see Tables 2 and 4). As

Table 4. **Results for Grounded Report Generation (GRG) on PadChest-GR and zero-shot VinDr-CXR.** We report mean IoU (↑), CheXbert F1 (micro/macro) (↑), CheXbert cosine similarity (↑), and CXRFEScore (↑). Bold values indicate the best score for each metric.

Model	PadChest-GR					VinDr-CXR (Zero-Shot)				
	IoU ↑	F1-Mi ↑	F1-Ma ↑	Cos. ↑	CXRFEScore ↑	IoU ↑	F1-Mi ↑	F1-Ma ↑	Cos. ↑	CXRFEScore ↑
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
CURE	**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

Table 5. **Results for Report Generation (RG) on the MIMIC-CXR test set.** We report CheXbert F1 (F1-Ma/Mi) (↑), Precision (P-Ma/Mi) (↑), and Recall (R-Ma/Mi) (↑), each macro (Ma) and micro (Mi) averaged together with CheXbert Cosine Similarity (Cos.) (↑), CXRFEScore (CXRFES) (↑), RaTEScore (RaTES) (↑), and RadGraph F1 (RadF1) (↑). **Bold** and underlined values indicate the best and second-best models per metric, respectively.

Model	F1-Ma ↑	F1-Mi ↑	P-Ma ↑	P-Mi ↑	R-Ma ↑	R-Mi ↑	Cos. ↑	CXRFES ↑	RaTES ↑	RadF1 ↑
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
CURE (GRG)	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)	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+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

shown in Table 2, CURE surpasses MAIRA-2 in grounding performance on the PG task, achieving higher Micro and Macro IoU. Similarly, on the GRG task, CURE equips MedGemma-4B-IT with a strong grounding ability, outperforming MAIRA-2 in IoU and narrowing the gap in text-based metrics, despite MAIRA-2's advantage from training on proprietary GRG-specific data. These results demonstrate that the capabilities learned through CURE generalize effectively to out-of-domain datasets and task formats.

Table 6. **Hallucination analysis for AGRG on a Chest ImaGenome subset.** We report abnormal finding hallucination rates (%) and assess report consistency using a Natural Language Inference (NLI) framework, reporting the fraction (%) of Contradiction (Cont.) and Entailment (Entail). Lower hallucination and contradiction rates, together with higher entailment rates, indicate better clinical faithfulness and grounding quality.

Model	Anatomy	Abn. ↓	Cont. ↓	Entail. ↑
MAIRA-2	Cardiac Silhouette	**2.00**	**8.00**	25.00
CURE	Cardiac Silhouette	25.67	27.67	**47.33**
MAIRA-2	Left Clavicle	59.00	22.67	5.00
CURE	Left Clavicle	**1.00**	**7.00**	**32.67**
MAIRA-2	Left Lung	12.00	56.33	21.67
CURE	Left Lung	**7.00**	**32.33**	**41.67**
MAIRA-2	Right Clavicle	62.67	20.33	1.67
CURE	Right Clavicle	**1.00**	**7.33**	**27.67**
MAIRA-2	Right Lung	**10.67**	53.67	29.33
CURE	Right Lung	11.67	**27.00**	**46.00**
MAIRA-2	Spine	12.67	38.33	13.00
CURE	Spine	**6.33**	**3.33**	**41.67**
MAIRA-2	*Mean Anatomies*	26.50	33.22	15.94
CURE	*Mean Anatomies*	**8.78**	**17.44**	**39.50**

4.2. Ablation Study

To assess the contribution of each component in our training pipeline, we conduct a detailed ablation study summarized in Table 7. For context, we include MAIRA-2 as a strong external baseline. Our analysis begins with a multi-task fine-tuned baseline (`v1`), trained without data augmentation, curriculum learning, or specialized pre-training.

Even this simplest variant already surpasses MAIRA-2 on the AGRG task and on the zero-shot VinDr-CXR PG task, highlighting the effectiveness of our fine-grained task reformulation. However, `v1` underperforms on GRG IoU and on the MS-CXR PG task. Adding **data augmentation** (`v2`) yields small but consistent improvements in PG performance across all three grounding datasets.

We then examine the effect of **curriculum learning** by varying the frequency of our performance-based reweighting strategy (`v3`–`v5`). A reweighting interval of 3000 steps (`v5`) tends to produce better performance, outperforming more frequent updates. We therefore adopt this interval in all subsequent experiments.

A larger gain emerges when introducing a dedicated **Chest ImaGenome pre-training stage** (`v6`–`v8`). Increasing the pre-training duration to 3000 steps (`v8`) enables the model to match MAIRA-2 on the MS-CXR PG task (0.495), emphasizing the value of grounding-specific pre-training before the curriculum stage.

Finally, we perform a hyperparameter search (HPS) for both the pre-training stage and the multi-task fine-tuning phase (`v9`). Our final model, **CURE (`v9`)**, achieves the best overall performance, notably surpassing MAIRA-2 on the challenging GRG (PC) IoU benchmark (0.265 vs. 0.256), while maintaining strong results across all other tasks.

Table 7. **Ablation Study.** We evaluate the contribution of each component across three grounding tasks. CXRS denotes the CXRFEScore metric. For Phrase Grounding (PG), we report Micro-Averaged IoU on MS-CXR (MS), PadChest-GR (PC), and VinDr-CXR (VD). Note that: CL(f) indicates curriculum learning with a reweighting frequency of f steps, CIG(s) denotes a Chest ImaGenome pre-training stage of s steps, HPS refers to hyperparameter search. **Bold** and underlined values indicate the best and second-best models per metric, respectively.

Model Configuration	AGRG (CIG)		GRG (PC)		GRG (VD)		PG (IoU Mi. ↑)		
	IoU ↑	CXRS ↑	IoU ↑	CXRS ↑	IoU ↑	CXRS ↑	MS	PC	VD
MAIRA-2 (External Baseline)	0.249 ± 0.008	0.357 ± 0.010	0.256 ± 0.011	**0.616** ± 0.011	0.217 ± 0.007	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)	0.380 ± 0.008	0.517 ± 0.011	0.171 ± 0.009	0.589 ± 0.011	0.207 ± 0.006	0.630 ± 0.007	0.388 ± 0.017	0.356 ± 0.007	0.191 ± 0.004
v2: + Aug	0.360 ± 0.009	0.522 ± 0.011	0.185 ± 0.010	0.599 ± 0.011	0.221 ± 0.007	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	0.521 ± 0.011	0.179 ± 0.010	0.592 ± 0.011	0.224 ± 0.007	0.630 ± 0.007	0.409 ± 0.019	0.383 ± 0.007	0.210 ± 0.005
v4: + Aug + CL(2k)	0.394 ± 0.009	0.513 ± 0.011	0.193 ± 0.010	0.596 ± 0.011	0.222 ± 0.006	0.651 ± 0.007	0.393 ± 0.017	0.383 ± 0.007	0.196 ± 0.005
v5: + Aug + CL(3k)	0.411 ± 0.009	0.526 ± 0.011	0.180 ± 0.010	0.595 ± 0.012	0.217 ± 0.007	**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	0.518 ± 0.011	0.195 ± 0.009	0.591 ± 0.011	0.232 ± 0.006	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	0.533 ± 0.011	0.203 ± 0.010	0.586 ± 0.011	0.227 ± 0.006	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	0.521 ± 0.011	0.207 ± 0.010	0.582 ± 0.011	0.233 ± 0.006	0.601 ± 0.007	0.495 ± 0.016	0.421 ± 0.007	0.224 ± 0.005
v9 (CURE): + Aug + CIG(3k) + CL(3k) + HPS	**0.601** ± 0.008	**0.549** ± 0.011	**0.265** ± 0.011	0.574 ± 0.010	**0.262** ± 0.007	0.540 ± 0.007	**0.552** ± 0.015	**0.453** ± 0.006	**0.243** ± 0.005

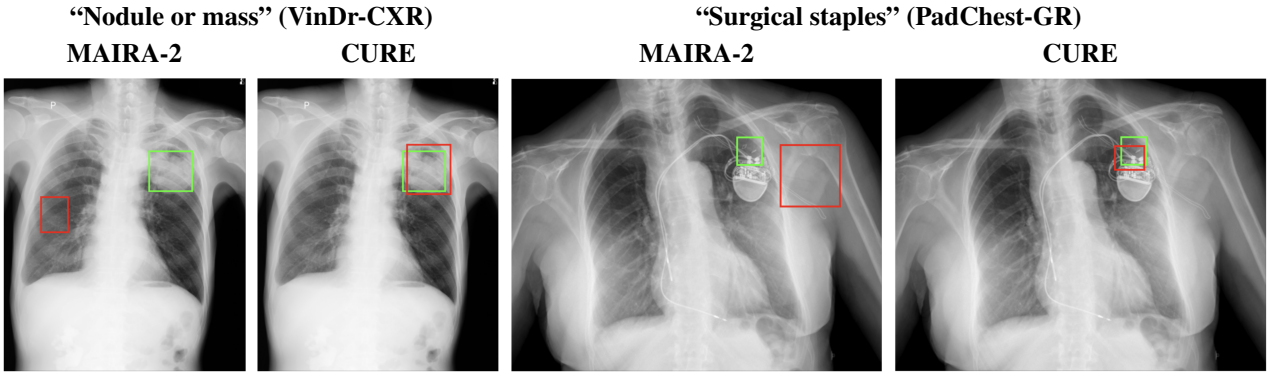


Figure 3. **Qualitative Examples.** Qualitative phrase grounding (PG) results on challenging examples from the VinDr-CXR and PadChest-GR datasets. The left panels show the detection of a *"Nodule or mass"* (VinDr-CXR), while the right panels demonstrate the grounding of *"Surgical staples"* (PadChest-GR). Ground-truth regions are shown in green for reference, and model predictions from **MAIRA-2** and **CURE** are shown in red.

4.3. Qualitative Analysis

Figure 3 compares the phrase grounding performance of CURE and MAIRA-2 on challenging examples from VinDr-CXR [32] and PadChest-GR [11]. In the left panel (VinDr-CXR), MAIRA-2 fails to localize a *"Nodule or mass"*, incorrectly placing a bounding box in the lower right lung field. CURE, conversely, accurately localizes the true nodule in the upper left lung, closely aligning with the ground truth. Similarly, in the right panel (PadChest-GR) for *"Surgical staples"*, MAIRA-2's prediction falls outside the relevant anatomical region, whereas CURE successfully bounds the staples and implanted device. These qualitative examples are consistent with CURE's overall higher phrase grounding metrics on both datasets, illustrating its improved spatial localization and semantic alignment compared to the baseline.

5. Conclusion

We introduced CURE, an error-aware curriculum learning framework that advances visual grounding and factual reliability in medical vision–language models through adaptive multi-task training on existing public datasets. By dynamically prioritizing underperforming samples and categories, CURE delivers consistent improvements over the baseline across diverse tasks, including phrase grounding, where it outperforms in all Micro and Macro IoU metrics on datasets such as MS-CXR, PadChest-GR, and VinDr-CXR. Moreover, CURE enables effective grounding in the base MedGemma-4B-IT model— which originally lacks visual grounding capabilities—achieving a +0.35 IoU improvement over MAIRA-2 in anatomy-grounded report generation, more than doubling localization accuracy. Finally, CURE reduces hallucinations by 18.6% across key anatomical regions, halves contradictions, and doubles entailment, demonstrating measurable progress in clinical faithfulness for medical VLMs.

Acknowledgments

This work was conducted while P. Messina was a remote research intern at the Image and Video Understanding Lab

(IVUL) at KAUST, under the supervision of B. Ghanem. P. Messina was supported by the ANID Scholarship Program (Doctorado Becas Chile 2019-21191569). We also acknowledge the support of Fondecyt grant 1231724. This work was also funded by ANID - Millennium Science Initiative Program - ICN2021_004 (iHEALTH) as well as ICN17_002 (IMFD), and by the National Center for Artificial Intelligence (CENIA) FB210017, Basal Funds for Centers of Excellence (ANID). The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. 2, 5

[3] Shruthi Bannur, Kenza Bouzid, Daniel Coelho de Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie Hyland. Maira-2: Grounded radiology report generation. Technical Report MSR-TR-2024-18, Microsoft, 2024. 1, 6, 22

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2

[5] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie Hyland, Harshita Sharma, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez Valle, Hoifung Poon, and Ozan Oktay. Ms-cxr: Making the most of text semantics to improve biomedical vision-language processing (version 1.1.0). *PhysioNet*, 2024. 3, 5

[6] Arnaud Boutillon, Pierre-Henri Conze, Christelle Pons, Valérie Burdin, and Bhushan Borotikar. Multi-task, multi-domain deep segmentation with shared representations and contrastive regularization for sparse pediatric datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–249. Springer, 2021. 2

[7] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 13

[8] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48, 1993. 2

[9] Chen Chen, Wenjia Bai, and Daniel Rueckert. Multi-task learning for left atrial segmentation on ge-mri. In *International workshop on statistical atlases and computational models of the heart*, pages 292–301. Springer, 2018. 2

[10] Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning. In *ICCVW*, 2025. 1

[11] Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120, 2025. 3, 5, 8

[12] Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12902–12915, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. 5, 16

[13] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[15] Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, 2024. 2

[16] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 2

[17] Alistair E. W. Johnson, Tom J. Pollard, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr database (version 2.1.0). *PhysioNet*, 2024. 5

[18] Béria Chingnabé Kalpélbé, Angel Gabriel Adaambiik, and Wei Peng. Vision language models in medicine. *arXiv preprint arXiv:2503.01863*, 2025. 2

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 7482–7491, 2018. 2

[20] Sangwook Kim, Thomas G Purdie, and Chris McIntosh. Cross-task attention network: Improving multi-task learning for medical imaging applications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–128. Springer, 2023. 2

[21] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010. 2

[22] Chaoyi Li, Meng Li, Can Peng, and Brian C Lovell. Dynamic curriculum learning via in-domain uncertainty for medical image classification. In *International conference on medical image computing and computer-assisted intervention*, pages 747–757. Springer, 2023. 2

[23] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 36:28541–28564, 2023. 1

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 1

[26] Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, and Chenguang Gui. Hierarchical prompt learning for multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10888–10898, 2023. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 5

[28] Pablo Messina, Rene Vidal, Denis Parra, Alvaro Soto, and Vladimir Araujo. Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3955–3986, Bangkok, Thailand, 2024. Association for Computational Linguistics. 5, 16

[29] Pablo Messina, René Vidal, Denis Parra, Álvaro Soto, and Vladimir Araujo. Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. *arXiv preprint arXiv:2407.01948*, 2024. 3

[30] Youssef Mohamed, Noran Mohamed, Khaled Abouhashad, Feilong Tang, Sara Atito, Shoaib Jameel, Imran Razzak, and Ahmed B Zaky. Deepchest: Dynamic gradient-free task weighting for effective multi-task learning in chest x-ray classification. *arXiv preprint arXiv:2505.23595*, 2025. 2

[31] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14788–14798, 2025. 2

[32] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020. 5, 8, 16

[33] Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. e-health CSIRO at RRG24: Entropy-augmented self-critical sequence training for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 99–104, Bangkok, Thailand, 2024. Association for Computational Linguistics. 6, 23, 24

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 12

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021. 1

[36] Maria V Sainz de Cea, Karl Diedrich, Ran Bakalo, Lior Ness, and David Richmond. Multi-task learning for detection and classification of cancer in screening mammography. In *International conference on medical image computing and computer-assisted intervention*, pages 241–250. Springer, 2020. 2

[37] Karen Sanchez, Carlos Hinojosa, Henry Arguello, Denis Kouamé, Olivier Meyrignac, and Adrian Basarab. Cx-dagan: Domain adaptation for pneumonia diagnosis on a small chest x-ray dataset. *IEEE Transactions on Medical Imaging*, 41 (11):3278–3288, 2022. 3

[38] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 1, 5

[39] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 2

[40] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 2

[41] Athira Kalladayil Shibu, Sriprabha Ramanarayanan, Vinoth Kanna, Jaikishan Jayakumar, Keerthi Ram, and Mohanasankar Sivaprakasam. Medsam-guided curriculum learning for white matter tract segmentation in block face imaging of fetal brain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1034–1041, 2025. 2

[42] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023. 2

[43] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. 2

[44] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, 2020. Association for Computational Linguistics. 5, 16

[45] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024. 1

[46] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1

[47] Andrés Villa, Juan León Alcázar, Motasem Alfarra, Vladimir Araujo, Alvaro Soto, and Bernard Ghanem. Eagle: Enhanced visual grounding minimizes hallucinations in instructional multimodal models. *arXiv preprint arXiv:2501.02699*, 2025. 1

[48] Andrés Villa, Juan Léon, Alvaro Soto, and Bernard Ghanem. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. In *CVPRW*, pages 492–502, 2025. 1

[49] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021. 2

[50] Leon Weninger, Qianyu Liu, and Dorit Merhof. Multi-task learning for brain tumor segmentation. In *International MICCAI brainlesion workshop*, pages 327–337. Springer, 2019. 2

[51] Maria Wimmer, Gert Sluiter, David Major, Dimitrios Lenis, Astrid Berg, Theresa Neubauer, and Katja Bühler. Multi-task fusion for improving mammography screening data classification. *IEEE Transactions on Medical Imaging*, 41(4):937–950, 2021. 2

[52] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 12

[53] Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest imagenome dataset (version 1.0.0). *PhysioNet*, 2021. 3, 5

[54] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 2

[55] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025. 1

[56] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34 (12):5586–5609, 2021. 2

[57] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA, 2024. Association for Computational Linguistics. 5, 16

[58] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. 1994. 5