TEXT-GUIDED ACTIVE IMAGE SELECTION FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training vision-language models (VLMs) typically relies on large-scale paired datasets. However, naïve data scaling is often suboptimal or even impractical when one modality is significantly more expensive to obtain than the other. In medical imaging, for example, text reports are considerably cheaper to transfer and store compared to their corresponding imaging scans. In this work, we introduce the task of *text-guided active image selection* to efficiently construct a dataset for training report generation VLMs. Unlike standard active learning, where the pool consists of unlabeled inputs, we instead formulate the selection task in the text space. We instantiate this formulation with a suite of acquisition functions that identify the most informative reports for the downstream image-to-text task; their corresponding images are then queried to expand the training set. Experiments on two large-scale chest X-ray datasets show that text-guided selection achieves performance comparable to image-text selection, where both modalities are available. Moreover, we show that acquisition functions improve over random selection in terms of CheXbert-F1 at small data budgets. Finally, we demonstrate the feasibility of text-guided selection for cross-institution acquisition. These findings motivate rethinking data selection for medical VLMs by shifting the acquisition process from the expensive image space to the more accessible text space.

1 INTRODUCTION

Recent progress in multi-modal AI has been driven by Vision-Language Models (VLMs) (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023) trained on large-scale collections of paired image-text data. Concurrent research has also identified empirical neural scaling laws (Kaplan et al., 2020; Hoffmann et al.; Cherti et al., 2023), demonstrating that the test loss decreases as a power-law function of training data, model size, and available compute. However, recent studies that have challenged this observation (Sorscher et al., 2022; Gadre et al., 2023) suggest that naïvely scaling uncurated datasets yields diminishing returns, while simple data pruning or curation methods improve data efficiency.

Importantly, data scaling and curation approaches implicitly assume that all modalities are available when the dataset is constructed. In real-world scenarios, however, this assumption might not hold as acquisition[1] costs, storage requirements, and regulatory constraints can vary significantly across modalities. This is particularly evident in the medical imaging domain. More specifically, imaging studies such as CT or MRI scans are typically large files stored on dedicated servers (PACS), and accessing them involves strict governance procedures due to privacy concerns (e.g., patient re-identification risks (Kaissis et al., 2020)). Thus, obtaining medical images can incur very high access and processing costs. In contrast, since radiology reports are generally easier to store and de-identify (Chambon et al., 2023), their acquisition cost is lower. As a result, researchers working within Trusted Research Environments (Kavianpour et al., 2022) may have immediate access to large collections of reports, whereas the acquisition of imaging studies at scale often incurs substantial retrieval costs and delays due to rigorous approval and extraction procedures.

This practical constraint raises the question: *If reports are readily available, while access to images is costly, how can we identify which images should be selected for training image-to-text models?*

[1]We use the terms "selection" and "acquisition" interchangeably, referring to data selection rather than imaging acquisition (i.e., performing a scan).

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

Table 1: Comparison of data selection strategies. ✔ denotes full data access, ● denotes partial access and ✘ denotes no access.

Strategy	Input x	Output y	Objective
Core-set Selection / Data Pruning	✔	✔	Minimize model development costs
Active Learning	✔	✘	
Partially Observable Active Learning / Sensing	●	✘	
Ours	✘	✔	

Motivated by this setting, we study the problem of text-guided active image selection for the vision-language task of radiology report generation. More specifically, given a large pool of candidate reports and a limited acquisition budget, our goal is to identify the subset of reports whose associated images should be retrieved next to maximize the model's downstream performance.

Unlike conventional active learning (Settles, 2009), where acquisition decisions are made in the input space (i.e., image space), our setting shifts the selection process to the output modality (i.e., text) dictated by the downstream task (cf. Figure 1). Consequently, by leveraging semantic signals in radiology reports, we aim to identify the most informative subset of image–report pairs to acquire for training VLMs.

The main **contributions** of our work are the following:

- We introduce a novel data acquisition paradigm based solely on text reports without accessing the corresponding images.
- We show empirically that text-only acquisition methods perform on par with their image-text counterparts.
- We demonstrate improvements over random selection at small acquisition budgets across diverse experimental settings.

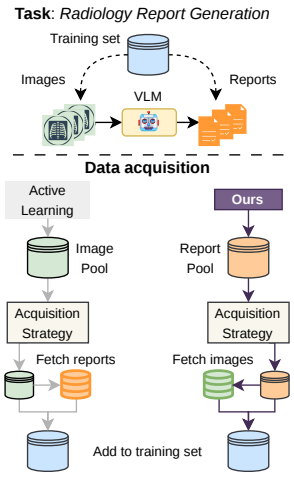


Figure 1: A high-level overview of the text-guided active image selection task.

2 Related Work

This work is informed by three key research areas. We begin by reviewing existing data selection strategies and highlight how their objectives differ from the task considered in this work (cf. Table 1). We then discuss the connection between our problem formulation and the theoretical framework of Bayesian optimal experimental design. Lastly, we discuss prior work on radiology report generation and highlight limitations of existing datasets, thus motivating the need for data selection in this task.

Core-set selection. A prominent approach in data-efficient learning is core-set selection (Mirzasoleiman et al., 2020), i.e., finding the most representative subset of the training data distribution. The field can be divided into training-free and training-based approaches. Training-free approaches aim to remove data points that are close to each other in the feature space (Sener & Savarese, 2017) or points where the model has high confidence (Coleman et al., 2020). Training-based approaches select samples that maximize the reduction in training loss (Paul et al., 2021), while other methods attempt to identify a core-set that best approximates the full dataset gradients (Mirzasoleiman et al., 2020; Killamsetty et al., 2021). Building on this line of work, recent data pruning methods can now scale to large datasets and modern VLMs (Wang et al., 2024; Xia et al., 2024; Yu et al., 2025). However, these methods assume access to both inputs and outputs at acquisition time which we want to steer away from in our proposed (and more challenging) setting (c.f. Table 1).

Active Learning. Another closely related area is pool-based active learning (Settles, 2009; Houlsby et al., 2011; Kirsch & Gal), which promotes data efficiency by querying the most informative samples from an unlabeled pool for annotation. In this direction, there exist uncertainty-based methods that use proxies such as the mutual information between predictions and model parameters (Houlsby

et al., 2011; Gal et al., 2017; Kirsch et al., 2019) or predictive uncertainty (Smith et al., 2023) for data selection. Moreover, connecting uncertainty with the core-set perspective discussed above, Ash et al. (2019) select data points that are simultaneously uncertain and diverse in the space of gradients. There also exist task-agnostic approaches (Sinha et al., 2019) that rely on adversarial training to identify unlabeled samples that are most dissimilar from the existing labeled set in the latent space. In the context of Large Language Models (LLMs), recent studies have also developed scalable active learning techniques for tasks such as code completion (Kobalczyk et al.), reward modeling (Shen et al., 2025) and supervised (Bhatt et al., 2024) or test-time (Hübotter et al.) fine-tuning. However, these methods assume access to a pool of unlabeled inputs (images), whereas our setting is constrained to a scenario where only downstream task outputs (text reports) are observable.

Partially Observable Active Learning & Active Feature Acquisition. In the context of partially observed data, methods focus on efficiently gathering features and/or labels to optimize a utility function, subject to an acquisition constraint. This leads to partial active learning Astorga et al. (2024), or active learning with missing values Han & Kang (2021), which utilizes feature imputation methods and uncertainty-based costs. While early work described this setup as active sensing Yu et al. (2009), the latter term has come to be synonymous with active classification Hollinger et al. (2016); Gao & Koller (2011), and, most importantly, active feature acquisition (AFA). AFA differs in that it assumes fully observed features at training time and shifts its focus to address the instance-level, sequential selection of missing features at inference time, where the goal is to optimize the trade-off between predictive performance and acquisition cost (Aronsson et al., 2025). Early approaches tackled this problem using model-based reinforcement learning (Li & Oliva, 2021), or discriminative models that estimate the conditional mutual information between unobserved features and the target label (Gadgil et al.). Recent work has focused on hybrid solutions such as learning feature dictionaries paired with policy networks (Huang et al., 2025), or imputing features with generative models followed by gradient-based acquisition (Norcliffe et al., 2025). This paradigm also extends to active modality acquisition (AMA) where the focus shifts on selecting entire data modalities rather than individual features. Prior work has explored learned policies (Kossen et al.), discriminative models that predict the expected value of future inputs given the observed history (Freyberg et al., 2024), or generative imputation techniques to optimize cohort-level (instead of sample-level) performance (Rheude et al., 2025). In contrast to AFA and AMA, our setting targets the efficient construction of a training dataset rather than improving performance at test time. To this end, we leverage the available outputs (reports) to guide data selection and, in turn, expand the training corpus. Moreover, we opt not to use policy-based or generative imputation approaches, which are designed for sequential decision-making under partial observability, as they would require training on an external database and may not transfer well to a given candidate pool due to distribution shifts (Kossen et al.).

Bayesian Optimal Experimental Design (BOED). BOED (Rainforth et al., 2024) provides a principled framework for selecting experiments that maximize the expected information gain (EIG) on model parameters $\boldsymbol{\theta}$, or the expected predictive information gain (EPIG) on the model's predictions over a held-out target set (Smith et al., 2023). Given that the estimation of EIG requires computationally expensive nested Monte Carlo (Rainforth et al., 2018), prior work has introduced variational approximations (Foster et al., 2019), implicit models that enable sampling without requiring closed-form densities (Kleinegesse & Gutmann, 2020), and likelihood-free policy networks that amortize the cost of BOED by predicting optimal designs from past observations (Ivanova et al., 2021). Recently, BOED has been applied to enhance the multi-turn conversation capabilities of LLMs by selecting optimal questions to pose to the user that maximize the EIG on the downstream task (Choudhury et al., 2025). In this work, we formulate the selection of informative reports through the lens of BOED, where a candidate report is defined as the design $\boldsymbol{y}$ and its corresponding image serves as the experimental outcome $\mathbf{X}$. A significant challenge in our setting is that the model learns the mapping $f_{\boldsymbol{\theta}} : \mathbf{X} \mapsto \boldsymbol{y}$, meaning that the inverse likelihood $p(\mathbf{X} \mid \boldsymbol{y}, \boldsymbol{\theta})$ required by BOED is not explicitly modeled. Therefore, our contribution lies in formulating acquisition functions over the text space while aiming for the same experimental outcome, i.e., obtaining the ground truth images corresponding to the selected subset of text reports and adding them to the training set.

Radiology Report Generation. The task of automated radiology report generation (Sloan et al., 2024), i.e., generating clinically accurate descriptions of findings from medical images, has seen rapid progress with the adoption of VLMs. Recent state-of-the-art approaches such as MAIRA-2 (Bannur

et al., 2024) and LLaVA-Rad (Zambrano Chaves et al., 2025) build on LLMs pre-trained on general-domain text data and adapt them to the radiology domain via parameter-efficient fine-tuning (Hu et al., 2021). Despite these advances, the task remains fundamentally constrained by data limitations. Large-scale datasets, e.g., MIMIC-CXR (Johnson et al., 2019) are heavily skewed towards normal studies, with pathological cases forming a long tail. Moreover, the fact that reports often contain references to prior, unavailable examinations (Ramesh et al., 2022; Zambrano Chaves et al., 2025) further degrades the quality of training data, leading to hallucinated findings in the generated reports. As a result, these issues highlight the importance of data-centric approaches that curate training data prior to model development.

3 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{X}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ denote a dataset consisting of n pairs of images $\mathbf{X}_i \in \mathcal{X}$ and their corresponding (tokenized) text reports $\boldsymbol{y}_i \in \mathcal{Y}$. Moreover, for image-to-text generation tasks, the VLM $f_{\boldsymbol{\theta}}$ learns an autoregressive conditional distribution over language tokens given an image $\mathbf{X}$ by minimizing the negative log-likelihood (NLL) of the ground truth report $\boldsymbol{y}$, i.e.,

$$\mathcal{L}(\mathbf{X}, \boldsymbol{y}; \boldsymbol{\theta}) = -\sum_{t=1}^{|\boldsymbol{y}|} \log p_{\boldsymbol{\theta}}(y_t \mid \mathbf{X}, \boldsymbol{y}_{<t}). \tag{1}$$

In our formulation, we assume that a small set of paired image-text data is initially drawn from $\mathcal{D}$ to warm-up the VLM $f_{\boldsymbol{\theta}}$, and is referred to as the *pre-training* set $\mathcal{D}_{pre} = \{(\mathbf{X}_j, \boldsymbol{y}_j)\}_{j=1}^{m}$, $m \ll n$. In addition, we define a disjoint held-out *validation* set $\mathcal{D}^* = \{(\mathbf{X}_k^*, \boldsymbol{y}_k^*)\}_{k=1}^{l}$, $l \ll n$, also drawn from $\mathcal{D}$. The *pool* of candidate text reports, from which we aim to select a batch of informative samples to expand the pre-training set $\mathcal{D}_{pre}$, is defined as $\mathcal{Y}_{cand} = \mathcal{Y} \setminus (\mathcal{Y}_{pre} \cup \mathcal{Y}^*)$. More formally, given a resource budget b, the goal is to identify a subset that maximizes the model's performance on the test set $\mathcal{D}_{\text{test}}$ in terms of some evaluation metric $\mathcal{M}$ (e.g., CheXbert-F1 score (Smit et al., 2020)), i.e.,

$$\mathcal{S}_{\text{opt}} = \underset{\mathcal{S} \subset \{1,\ldots,|\mathcal{Y}_{cand}|\}}{\arg\max} \; \mathcal{M}\left(f_{\boldsymbol{\theta}'_{\mathcal{S}}}, \mathcal{D}_{\text{test}}\right) \quad \text{s.t.} \quad \sum_{s \in \mathcal{S}} \kappa_s \leq b, \tag{2}$$

where $f_{\boldsymbol{\theta}'_{\mathcal{S}}}$ refers to the model trained on both $\mathcal{D}_{\text{pre}}$ and the acquired subset after querying the database for the ground truth images corresponding to the selected reports, i.e., $\mathcal{D}_{\mathcal{S}} = \{(\mathbf{X}_s, \boldsymbol{y}_s)\}_{s \in \mathcal{S}}$, while κ_s denotes the cost of acquiring the s-th sample. In our study, we assume uniform acquisition costs across candidate samples, i.e., $\kappa_s = 1, \forall s \in \mathcal{S}$, and simplify the constraint to a fixed acquisition batch size of $|\mathcal{S}| = b$. Furthermore, since optimizing for the objective defined in Equation 2 is intractable in practice, the selection mechanism is implemented by a *model-based acquisition function* $\alpha(\boldsymbol{y}_s; \boldsymbol{\theta})$ that assigns a score to each candidate report $\boldsymbol{y}_s \in \mathcal{Y}_{\text{cand}}$ based on an informativeness criterion (e.g., reduction in predictive uncertainty). In that sense, as higher acquisition scores correspond to more informative data, the decision rule for selecting candidates is defined as:

$$\hat{\mathcal{S}}_{\text{opt}} = \underset{\substack{\mathcal{S} \subset \{1,\ldots,|\mathcal{Y}_{cand}|\} \\ |\mathcal{S}| = b}}{\arg\max} \sum_{s \in \mathcal{S}} \alpha(\boldsymbol{y}_s; \boldsymbol{\theta}). \tag{3}$$

3.1 Information-Theoretic Perspective

To motivate the task of text-guided active image selection, we ask whether acquisition decisions made in the text space can approximate those made in the image space, as in standard active learning, provided that the reports are sufficiently informative about the underlying visual content. This reduces to the question: *For image-to-text tasks, can data selection be performed in the output (text) space instead of the input (image) space without introducing noise into the selection process*? We begin by formalizing the intuition that more informative text reports reduce uncertainty over the image space.

Proposition 1 (Reduction in Image Uncertainty)**.** *Let $\mathbf{X}$ denote an image, and let $\mathbf{y}_1, \mathbf{y}_2$ be two text descriptions of $\mathbf{X}$. Assuming that $\mathbf{X} \to \mathbf{y}_1 \to \mathbf{y}_2$, which implies that $\mathbf{y}_1$ provides a more fine-grained description of $\mathbf{X}$ compared to $\mathbf{y}_2$, we have: $H(\mathbf{X} \mid \mathbf{y}_1) \leq H(\mathbf{X} \mid \mathbf{y}_2)$.*

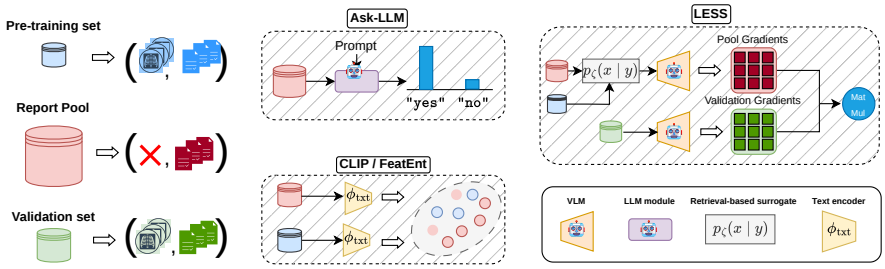


Figure 2: Illustration of model-based acquisition functions.

Proof. This follows directly from the Data Processing Inequality:

$$I(\mathbf{X};\mathbf{y}_1) \geq I(\mathbf{X};\mathbf{y}_2) \Leftrightarrow H(\mathbf{X}) - H(\mathbf{X} \mid \mathbf{y}_1) \geq H(\mathbf{X}) - H(\mathbf{X} \mid \mathbf{y}_2)$$
$$\Leftrightarrow H(\mathbf{X} \mid \mathbf{y}_1) \leq H(\mathbf{X} \mid \mathbf{y}_2).$$

□

Proposition 1 states that more informative reports restrict the space of compatible images, thus reducing selection noise by minimizing the uncertainty in the "inverse mapping" from text to image space. This suggests that text-space selection can serve as a viable proxy for multi-modal selection, with selection fidelity depending on how well text descriptions reflect the underlying visual content.

Empirical validation using BiomedCLIP. Since $H(\mathbf{X} \mid \mathbf{y})$ is intractable to estimate given an image-to-text VLM $f_{\boldsymbol{\theta}}$, we introduce an empirical proxy using the pre-trained BiomedCLIP model (Zhang et al., 2023), which provides a shared multi-modal latent space independent of $\boldsymbol{\theta}$. We evaluate whether more informative text induces a more concentrated set of semantically consistent images in this space. To this end, using a pool of images and radiology reports from MIMIC-CXR (Johnson et al., 2019), while controlling for pathology (*Pleural Effusion*) and patient gender (*Female*) to mitigate confounding effects, we construct two types of text queries:

(i) the ground truth radiology report $\boldsymbol{y}_1$;
(ii) the corresponding coarse pathology label $\boldsymbol{y}_2$.

Table 2: Effective Rank (ER) of retrieved image embeddings per type of text query.

Text query	ER
Report y_1	**29.06**
Label y_2	511.99

For each text query, we retrieve the most similar image from the pool using cosine similarity in the BiomedCLIP latent space. This yields a set of retrieved image embeddings per query type. We then compute the effective rank (Roy & Vetterli, 2007) of these embeddings to quantify their spread. As shown in Table 2, detailed radiology reports yield substantially lower effective rank than pathology labels, indicating that they map to a more concentrated region of the image embedding space. This observation is consistent with Proposition 1 and supports the use of text reports to guide data selection when images are not directly accessible.

4 Acquisition Functions

Building on the data selection framework defined in the previous section, we now instantiate it with a suite of acquisition functions designed to capture diverse notions of informativeness (Figure 2). More specifically, we quantify the utility of a candidate report through: (i) gradient-based estimation of influence on held-out data (`LESS`, Section 4.1); (ii) an LLM-based scoring method using the VLM's language model (`Ask-LLM`, Section 4.2); and (iii) feature space coverage (`FeatEnt`, Section 4.3, and `CLIP`, Section 4.4). In the following, we will denote a candidate text report as $\boldsymbol{y}_{\text{cand}}$.

4.1 Gradient-based Influence Estimation

This method is largely inspired by the influence estimation technique of Xia et al. (2024) which relies on a first-order approximation of training dynamics (Pruthi et al., 2020) to identify the candidate samples that will lead to the largest reduction in loss $\mathcal{L}(\cdot)$ (Equation 1) on a validation set

$\mathcal{D}^*$. More specifically, suppose that the VLM $f_{\boldsymbol{\theta}}$ has been trained for n_e epochs with the Adam optimizer (Kingma & Ba, 2014), η_i is the learning rate at epoch $i = 1, \ldots, n_e$, and $\boldsymbol{\theta}_i$ are model weights after the i-th epoch. The influence of a data point $\boldsymbol{z} = (\mathbf{X}, \boldsymbol{y})$ on the loss of a validation data point $\boldsymbol{z}' = (\mathbf{X}', \boldsymbol{y}')$ over the entire training run can be approximated as follows (Xia et al., 2024):

$$\text{Inf}_{\text{Adam}}(\boldsymbol{z}, \boldsymbol{z}') \approx \sum_{i=1}^{n_e} \eta_i \langle \nabla \mathcal{L}(\boldsymbol{z}'; \boldsymbol{\theta}_i), \boldsymbol{\Gamma}(\boldsymbol{z}; \boldsymbol{\theta}_i) \rangle,$$

where $\boldsymbol{\Gamma}(\boldsymbol{z}, \boldsymbol{\theta}_i)$ refers to Adam gradients for the training sample $\boldsymbol{z}$, while $\nabla \mathcal{L}(\boldsymbol{z}'; \boldsymbol{\theta}_i)$ denotes SGD gradients for the validation sample $\boldsymbol{z}'$. In practice, the gradients are first extracted from selected trainable layers of $f_{\boldsymbol{\theta}}$ (e.g., LoRA layers (Hu et al., 2021)) and then projected onto a lower-dimensional space using random projections (see Section 4.1 in Xia et al. (2024) for more information).

However, in order to utilize this method in our problem setup where we only observe a pool of text reports, we need to first construct approximate, yet semantically similar, image-text pairs. To this end, we define a retrieval-based surrogate that maps each candidate report $\boldsymbol{y}_{\text{cand}}$ to a proxy image $\tilde{\mathbf{X}}$ from the pre-training set $\mathcal{D}_{\text{pre}}$ that shares the same ground truth pathology labels. This allows us to approximate the training loss for a given candidate sample.

Formulation 1 (Retrieval-based Surrogate)**.** *Let $L : \mathcal{X} \cup \mathcal{Y} \to \mathcal{P}(\mathcal{V})$ denote the mapping from data points to a subset of the label set $\mathcal{V}$ (e.g., via a text-based classifier). To form approximate $(\tilde{\mathbf{X}}, \boldsymbol{y}_{cand})$ pairs for each text report in the pool $\mathcal{Y}_{cand}$, we select the image $\tilde{\mathbf{X}}$ from the pre-training set $\mathcal{D}_{pre}$ whose associated label set maximizes the Jaccard similarity with the ground truth label set of $\boldsymbol{y}_{cand}$, i.e.,*

$$\tilde{\mathbf{X}} = \underset{\mathbf{X} \in \mathcal{D}_{pre}}{\arg\max} \left(\frac{|L(\boldsymbol{y}_{cand}) \cap L(\mathbf{X})|}{|L(\boldsymbol{y}_{cand}) \cup L(\mathbf{X})|} \right).$$

Thus, we define the empirical surrogate $p_\zeta(\mathbf{X} \mid \boldsymbol{y}_{cand})$ as a Kronecker delta: $p_\zeta(\mathbf{X} \mid \boldsymbol{y}_{cand}) = \delta(\mathbf{X} - \tilde{\mathbf{X}})$.

This allows us to define the `LESS` acquisition function as:

$$\alpha_{\texttt{LESS}}(\boldsymbol{y}_{\text{cand}}) = \sum_{i=1}^{n_e} \eta_i \frac{\langle \bar{\nabla} \mathcal{L}(\mathcal{D}^*; \boldsymbol{\theta}_i), \hat{\boldsymbol{\Gamma}}(\tilde{\mathbf{X}}, \boldsymbol{y}_{\text{cand}}; \boldsymbol{\theta}_i) \rangle}{\| \langle \bar{\nabla} \mathcal{L}(\mathcal{D}^*; \boldsymbol{\theta}_i) \| \| \hat{\boldsymbol{\Gamma}}(\tilde{\mathbf{X}}, \boldsymbol{y}_{\text{cand}}; \boldsymbol{\theta}_i) \|}, \tag{4}$$

where $\bar{\nabla} \mathcal{L}(\mathcal{D}^*; \boldsymbol{\theta}_i) = \frac{1}{|\mathcal{D}^*|} \sum_{(\mathbf{X}^*, \boldsymbol{y}^*) \in \mathcal{D}^*} \hat{\nabla} \mathcal{L}(\mathbf{X}^*, \boldsymbol{y}^*; \boldsymbol{\theta}_i)$, while both $\hat{\nabla}$ and $\hat{\boldsymbol{\Gamma}}$ denote gradients from selected layers of the VLM after random projection (see Appendix C for more details).

4.2 LLM-BASED SCORING

Inspired by the recent work of Sachdeva et al. (2024) that uses LLMs for data curation, we propose a method that prompts the VLM's underlying language model with a candidate report $\boldsymbol{y}_{\text{cand}}$ and an instruction prompt that specifies the data selection task (see Figure 8 in the Appendix). Based on the model's softmax probabilities for the tokens `"yes"` and `"no"`, the acquisition function is defined as:

$$\alpha_{\texttt{Ask-LLM}}(\boldsymbol{y}_{\text{cand}}) = \frac{p_\theta(\texttt{"yes"} \mid \text{prompt}, \boldsymbol{y}_{\text{cand}})}{p_\theta(\texttt{"yes"} \mid \text{prompt}, \boldsymbol{y}_{\text{cand}}) + p_\theta(\texttt{"no"} \mid \text{prompt}, \boldsymbol{y}_{\text{cand}})}. \tag{5}$$

4.3 MODEL-BASED FEATURE SPACE DIVERSITY (FEATENT)

To encourage training set diversity, here we prioritize acquisitions that maximize coverage in the model's latent space. To this end, we first represent each candidate sample by its last-layer embedding extracted from the underlying LLM. In turn, inspired by Melo et al. (2024), we use a non-parametric, particle-based entropy estimator (Kraskov et al., 2004) based on k-nearest-neighbour (kNN) distances to compute the coverage of the existing training set if we were to acquire $\boldsymbol{y}_{\text{cand}}$. The `FeatEnt` acquisition function is defined as follows (note that $\psi(\cdot)$ is the digamma function):

$$\alpha_{\texttt{FeatEnt}}(\boldsymbol{y}_{\text{cand}}) = \log D(\boldsymbol{y}_{\text{cand}}) - \frac{1}{d_{\text{LLM}}} \psi\big(n_Y(\boldsymbol{y}_{\text{cand}}) + 1\big), \tag{6}$$

where d_{LLM} is the feature dimensionality, $D(\boldsymbol{y}_{\text{cand}})$ is the distance from the candidate to its k-th nearest neighbor in the pool, and $n_Y(\boldsymbol{y}_{\text{cand}}) = |\{\boldsymbol{y} \in \mathcal{Y}_{\text{pre}} \cup \mathcal{Y}_{\text{acq}} : \|\boldsymbol{y} - \boldsymbol{y}_{\text{cand}}\| \leq D(\boldsymbol{y}_{\text{cand}})\}|$ is the number of the pre-training $\mathcal{Y}_{\text{pre}}$ and the already acquired data points $\mathcal{Y}_{\text{acq}}$ within that neighbourhood.

4.4 Surrogate-based Feature Space Coverage (CLIP)

Instead of relying on the model's latent representations, this acquisition function measures coverage in the embedding space of an external CLIP model (Radford et al., 2021). More specifically, we embed the candidate report $\boldsymbol{y}_{\text{cand}}$ and each report in the pre-training set, $\boldsymbol{y}_{\text{pre}} \in \mathcal{D}_{\text{pre}}$, with the CLIP text encoder $\phi_{\text{txt}}(\cdot)$. We then define the CLIP acquisition function as the minimum cosine distance between the candidate and the pre-training reports:

$$\alpha_{\texttt{CLIP}}(\boldsymbol{y}_{\text{cand}}) = 1 - \max_{\boldsymbol{y}_{\text{pre}} \in \mathcal{D}_{\text{pre}}} \left(\frac{\phi_{\text{txt}}(\boldsymbol{y}_{\text{cand}})^\top \phi_{\text{txt}}(\boldsymbol{y}_{\text{pre}})}{\|\phi_{\text{txt}}(\boldsymbol{y}_{\text{cand}})\| \|\phi_{\text{txt}}(\boldsymbol{y}_{\text{pre}})\|} \right). \tag{7}$$

5 Empirical Setup

Datasets. We conduct experiments on MIMIC-CXR (Johnson et al., 2019) using LLaVA-Rad annotations (Zambrano Chaves et al., 2025) and CheXpert-Plus (Chambon et al., 2024), two large-scale multi-label datasets of chest X-rays accompanied by radiology reports. For all experiments, we consider only frontal chest X-rays. Additionally, we exclude reports from CheXpert-Plus that either lack an Impression section or contain fewer than 51 characters in the Findings section. For our study, we construct three subsets per dataset. First, we randomly sample 1,000 image-text pairs using an iterative greedy strategy to construct a pathology-stratified pre-training set $\mathcal{D}_{\text{pre}}$. Second, following the same procedure, we construct a validation set $\mathcal{D}^*$ with 1,000 image-text pairs. Last, we sample radiology reports to form the pool $\mathcal{Y}_{\text{cand}}$, corresponding to 25% (99,510 samples) of the official MIMIC-CXR training split and 100% (45,734 samples) of the CheXpert-Plus training set, while preserving the original pathology distribution. Note that we repeat experiments across three random seeds, each yielding a different data split, and report the mean and standard deviation of all metrics. In terms of downstream evaluation, we report results on the official test split of each dataset.

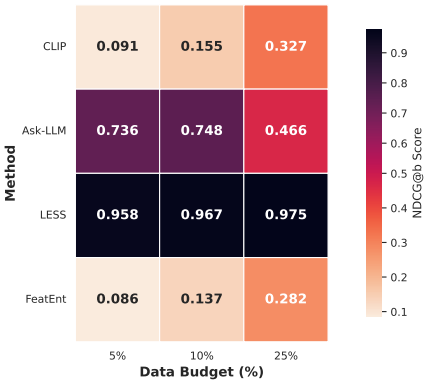
For the cross-institution acquisition experiments, we consider the settings where the pre-training, validation, and test sets come from one institution, i.e., either CheXpert-Plus or IU-Xray (Demner-Fushman et al., 2016) ($|\mathcal{D}_{\text{pre}}| = 1,961$ and $|\mathcal{D}^*| = 280$), while the pool is drawn from another institution (MIMIC-CXR).

Models. Following the design of state-of-the-art medical VLMs (Bannur et al., 2024; Zambrano Chaves et al., 2025), we adopt `vicuna-7b-v1.5` (Zheng et al., 2023) as the backbone and apply LoRA fine-tuning (Hu et al., 2021) to its layers. Visual features are extracted using the BiomedCLIP (Zhang et al., 2023) vision encoder, which was pre-trained on a separate dataset (PMC-15M), and are projected onto the LLM's embedding space via a multi-modal projector trained from scratch. Full training details and hyper-parameters are provided in Appendix B.

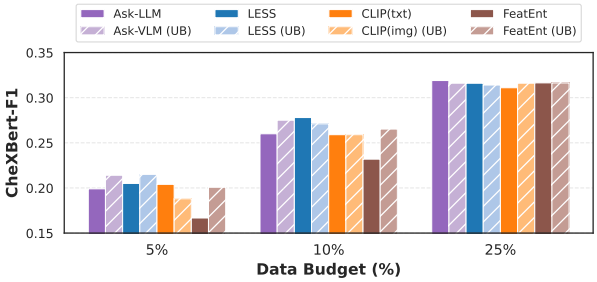
Evaluation. Following standard practice (Zambrano Chaves et al., 2025), we evaluate model performance on the task of radiology report generation using both lexical and clinical metrics. Specifically, we assess lexical similarity with reference reports using the metrics BLEU-1–4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004). To assess clinical validity, we report CheXBert-F1 (Smit et al., 2020) over 14 pathologies, as well as RadGraph-F1 (Jain et al.) to measure the accuracy of extracted clinical entities and their relations. Moreover, since lexical metrics correlate weakly with clinical correctness (Sloan et al., 2024), we prioritize clinical metrics in our analysis below. Full results are provided in Appendix D.

6 Experiments

For our experiments, we consider the following acquisition methods (described in Section 4): `LESS`, `Ask-LLM`, `FeatEnt`, and `CLIP`, along with the `Random` acquisition baseline. We compare these methods in terms of downstream report generation performance on the official test sets across different



(a) Ranking agreement between text-only and image-text acquisition using NDCG@b (↑).



(b) Downstream performance (CheXbert-F1). (UB) denotes the upper-bound (image-text) variant of an acquisition method.

Figure 3: Comparison of *text-only* and *image-text* acquisition. Ideally, text-only methods should select similar samples and achieve comparable performance to their upper-bound, image-text counterparts.

acquisition budgets, measured as percentages of the available candidate data (e.g., 5% of the pool). Additional experiments and analysis are provided in Appendices E and F.

6.1 Text-only vs. Multi-modal acquisition

To assess whether text-only data selection is a reliable proxy in the absence of ground truth candidate images, we compare each acquisition function against its upper-bound counterpart, which has access to both candidate images and reports, or just images in the case of `CLIP` (see Appendix C.2 and C.3 for `Ask-LLM` and `CLIP` upper-bound definitions). Specifically, we assess whether text-only methods (i) select similar subsets under a given budget b, and (ii) preserve the ranking of candidate samples induced by their upper-bound counterparts. We measure ranking agreement using Normalized Discounted Cumulative Gain (NDCG@b), treating each upper-bound ranking as the reference, and compare downstream performance by training a model on the subset selected by each approach.

Overall, Figure 3 shows that, in our experiments, text-only acquisition achieves downstream performance comparable to image-text acquisition. `LESS` selects subsets with over 95% overlap with its upper-bound counterpart, while `Ask-LLM` closely agrees with `Ask-VLM` and performs within $\sim 2\%$ in CheXbert-F1 at 5% and 10% budgets. Although `CLIP(txt)` selects substantially different subsets from `CLIP(img)`, both achieve similar downstream performance. `FeatEnt` also achieves $\sim 3\%$ lower CheXbert-F1 than its upper-bound variant at small data budgets.

6.2 Acquisition results

We report acquisition results on MIMIC-CXR and CheXpert-Plus in Table 3. Note that, for all acquisition methods considered here, we use pathology-stratified acquisition on MIMIC-CXR since it further improves downstream performance (see Appendix F.2). For CheXpert-Plus, however, we use naïve top-k selection for all methods, as pathology stratification did not improve performance. Overall, `LESS` outperforms random selection in terms of CheXbert-F1 at 5% and 10% data budgets on MIMIC-CXR, and at the 10% budget on CheXpert-Plus, while `FeatEnt` achieves the highest CheXbert-F1 at 5% budget on CheXpert-Plus (see Appendix D for full results).

Cross-institution acquisition. We further evaluate the ability of our acquisition methods to identify high-utility data subsets from an external database. Results are presented in Table 4. In general, acquiring more data from the external database (MIMIC-CXR) improves CheXbert-F1 on the source domain test set (i.e., either CheXpert-Plus or IU-Xray). Notably, `Ask-LLM` outperforms random selection at 5% and 25% data budgets in the CheXpert-Plus/MIMIC-CXR scenario. However, in the IU-Xray/MIMIC-CXR setting, performance differences between acquisition methods, including the random baseline, are modest, which highlight the difficulty of the task.

Table 3: Report generation results across acquisition methods and data budgets. †*Upper*: upper-bound reference setting for image-text selection, included for comparison but excluded from the ranking.

Dataset	Method	5%			10%			25%		
		BLEU-1	CheXbert-F1-14		BLEU-1	CheXbert-F1-14		BLEU-1	CheXbert-F1-14	
			Macro ↑	Micro ↑		Macro ↑	Micro ↑		Macro ↑	Micro ↑
MIMIC	*Upper*†	–	0.214	0.298	–	0.276	0.394	–	0.317	0.447
	Random	26.3 ±0.8	0.199 ±0.012	0.277 ±0.021	30.0 ±0.2	0.265 ±0.006	0.372 ±0.006	30.9 ±0.0	0.315 ±0.001	**0.444** ±0.002
	LESS	29.7 ±1.2	**0.219** ±0.027	**0.320** ±0.041	31.1 ±1.4	**0.271** ±0.009	**0.384** ±0.015	31.4 ±1.6	0.313 ±0.005	0.441 ±0.004
	FeatEnt	27.0 ±1.6	0.212 ±0.013	0.297 ±0.017	30.3 ±0.7	0.253 ±0.015	0.361 ±0.015	31.4 ±1.6	0.313 ±0.005	0.441 ±0.004
	Ask-LLM	**31.9** ±0.6	0.207 ±0.006	0.287 ±0.009	**33.0** ±0.4	0.269 ±0.009	0.373 ±0.006	**32.5** ±0.1	**0.316** ±0.003	0.442 ±0.003
	CLIP	26.5 ±0.8	0.202 ±0.013	0.275 ±0.018	29.5 ±0.3	0.258 ±0.001	0.354 ±0.002	30.3 ±0.3	0.311 ±0.002	0.439 ±0.004
CheXpert	*Upper*†	–	0.229	0.362	–	0.259	0.426	–	0.258	0.450
	Random	20.4 ±0.8	0.203 ±0.001	0.342 ±0.013	24.9 ±0.3	0.222 ±0.005	0.360 ±0.010	27.2 ±0.9	0.271 ±0.005	0.437 ±0.006
	LESS	21.2 ±1.2	0.163 ±0.012	0.280 ±0.032	**27.5** ±2.8	**0.243** ±0.003	**0.431** ±0.008	28.3 ±1.5	0.272 ±0.013	0.437 ±0.010
	FeatEnt	19.7 ±1.0	**0.220** ±0.012	0.370 ±0.022	24.3 ±1.6	0.226 ±0.008	0.390 ±0.007	26.8 ±1.7	0.269 ±0.008	**0.459** ±0.004
	Ask-LLM	15.2 ±0.6	0.190 ±0.022	**0.371** ±0.018	18.6 ±1.2	0.224 ±0.012	0.409 ±0.027	**28.8** ±0.3	**0.274** ±0.010	0.453 ±0.005
	CLIP	**21.4** ±1.2	0.190 ±0.009	0.334 ±0.015	23.1 ±0.9	0.222 ±0.012	0.381 ±0.014	24.8 ±0.5	0.251 ±0.008	0.433 ±0.005

Table 4: Report generation results across acquisition methods and data budgets for the *cross-institution acquisition setting*.

Dataset (Pre-train & Test/Pool)	Method	5%			10%			25%		
		BLEU-1	CheXbert-F1-14		BLEU-1	CheXbert-F1-14		BLEU-1	CheXbert-F1-14	
			Macro ↑	Micro ↑		Macro ↑	Micro ↑		Macro ↑	Micro ↑
CheXpert/MIMIC-CXR	Random	20.4	0.200	0.339	19.3	**0.233**	**0.389**	21.6	0.245	0.403
	LESS	**24.8**	0.213	0.347	20.5	0.203	0.346	22.4	0.241	0.377
	FeatEnt	19.5	0.200	**0.360**	19.7	0.230	**0.389**	**23.5**	0.254	**0.418**
	Ask-LLM	21.9	**0.220**	0.350	**24.1**	0.230	0.376	23.1	**0.273**	0.417
IU-Xray/MIMIC-CXR	Random	**38.8**	0.070	0.526	**39.7**	**0.123**	**0.565**	38.7	**0.167**	0.570
	LESS	36.9	0.072	0.530	37.8	0.107	0.557	37.7	0.151	**0.571**
	FeatEnt	36.1	0.056	**0.544**	38.5	0.091	0.550	**40.5**	0.133	0.561
	Ask-LLM	35.7	**0.078**	0.540	38.1	0.089	0.560	39.9	0.159	0.570

7 Conclusion

In this work, we introduce the task of *text-guided active image selection* for efficiently constructing training datasets for report generation VLMs. We realize this framework using a suite of acquisition functions to identify a subset of reports whose corresponding images should be queried and added to the training set. We empirically show that our approach closely matches the performance of image-text selection, while avoiding the cost of processing candidate images at selection time. We also show that text-guided selection yields high-utility data subsets at small budgets across both within- and cross-institution settings. These results suggest that VLM development costs can be significantly reduced through principled data selection, even in cases where only reports are available.

Limitations and Future Work. The proposed framework depends on the extent to which the available reports capture the variability present in their corresponding images, which are not observed during data selection. For example, two similar candidate reports may correspond to images whose visual characteristics differ significantly, and those differences are not reflected in text. As a result, these candidates will receive similar acquisition scores based only on the available text information. This limitation may arise in cases where reports are incomplete or highly templated. Moreover, improvements over random selection are not consistent across evaluation metrics. While we observe notable gains in CheXbert-F1, the differences are less pronounced for RadGraph-F1 (Appendix D). We also adopt a standard VLM fine-tuning approach in which only the multi-modal projector and LoRA layers are trained, while the vision encoder remains frozen; however, it is worth exploring whether adapting the vision encoder after data selection yields additional gains.

AI USE STATEMENT

During the preparation of this work, the authors used Large Language Models (LLMs) solely as writing assistants to refine sentence structure, improve linguistic clarity, and assist with identifying relevant literature. All core intellectual contributions, including the formulation and refinement of research hypotheses, the conceptualization of experimental designs, formal mathematical derivations, and the analysis and interpretation of empirical results, were conducted by the authors. The authors retain full accountability for the content, accuracy, and integrity of the final manuscript.

ETHICS STATEMENT

This work introduces a framework for text-guided active image selection that reduces the amount of image-text data required to train report generation VLMs. The proposed approach enables data selection using text reports without requiring access to the corresponding images during the selection stage. This can reduce the need to retrieve, store, or share sensitive medical imaging data at scale. Note, however, that radiology reports may also contain sensitive information, and therefore might require appropriate privacy and access controls.

We acknowledge several potential ethical risks. Data selection may introduce or amplify biases present in the underlying datasets, including biases related to patient demographics, disease prevalence, or radiologists' reporting practices. Such biases may affect the resulting models' performance and fairness across different patient populations. These risks are particularly relevant in medical applications, where generated reports may contain hallucinated or incorrect clinical findings and could potentially lead to harmful decisions if used without appropriate clinical oversight.

The models developed in this work have not been extensively validated across diverse clinical populations and deployment settings. Therefore, they are not intended for clinical use or decision-making. We use existing medical imaging datasets in accordance with their respective access and usage requirements and do not collect new patient data as part of this work.

REPRODUCIBILITY STATEMENT

Throughout this work, we provide detailed information to facilitate reproducibility of our experimental setup. Dataset details are provided in Section 5, with Appendix A describing the construction of the various data splits. For fine-tuning, we follow standard practices established by state-of-the-art medical VLMs (Bannur et al., 2024; Zambrano Chaves et al., 2025). Full details on specific model choices are provided in Section 5. We summarize the fine-tuning hyper-parameters in Appendix B and provide a detailed description of the acquisition functions, including their design choices and hyper-parameters, in Section 4 and Appendix C.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Linus Aronsson, Arman Rahbar, and Morteza Haghir Chehreghani. A survey on active feature acquisition strategies. *arXiv preprint arXiv:2502.11067*, 2025.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Nicolás Astorga, Tennison Liu, Nabeel Seedat, and Mihaela van der Schaar. Active learning with llms for partially observed and cost-aware scenarios. *Advances in Neural Information Processing Systems*, 37:20819–20857, 2024.

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.

Gantavya Bhatt, Yifang Chen, Arnav Das, Jifan Zhang, Sang Truong, Stephen Mussmann, Yinglun Zhu, Jeff Bilmes, Simon Du, Kevin Jamieson, et al. An experimental design framework for label-efficient supervised finetuning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6549–6560, 2024.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Pierre J Chambon, Christopher Wu, Jackson M Steinkamp, Jason Adleberg, Tessa S Cook, and Curtis P Langlotz. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *Journal of the American Medical Informatics Association*, 30(2):318–328, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. Bed-llm: Intelligent information gathering with llms and bayesian experimental design. *arXiv preprint arXiv:2508.21184*, 2025.

Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin*, 114(3):494, 1993.

C Coleman, C Yeh, S Mussmann, B Mirzasoleiman, P Bailis, P Liang, J Leskovec, and M Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. *Advances in neural information processing systems*, 32, 2019.

Jan Freyberg, Abhijit Guha Roy, Terry Spitz, Beverly Freeman, Mike Schaekermann, Patricia Strachan, Eva Schnider, Renee Wong, Dale R Webster, Alan Karthikesalingam, et al. Mint: a wrapper to make multi-modal and multi-image ai models interactive. *arXiv preprint arXiv:2401.12032*, 2024.

Soham Gadgil, Ian Connick Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *The Twelfth International Conference on Learning Representations*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Tianshi Gao and Daphne Koller. Active classification based on value of classifier. *Advances in neural information processing systems*, 24, 2011.

Jongmin Han and Seokho Kang. Active learning with missing values considering imputation uncertainty. *Knowledge-Based Systems*, 224:107079, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models.

Geoffrey A Hollinger, Urbashi Mitra, and Gaurav S Sukhatme. Active classification: Theory and application to underwater inspection. In *Robotics Research: The 15th International Symposium ISRR*, pp. 95–110. Springer, 2016.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hung-Tien Huang, Dzung Dinh, and Junier B Oliva. Information templates: A new paradigm for intelligent active feature acquisition. *arXiv preprint arXiv:2508.18380*, 2025.

Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. Efficiently learning at test-time: Active fine-tuning of llms. In *The Thirteenth International Conference on Learning Representations*.

Desi R Ivanova, Adam Foster, Steven Kleinegesse, Michael U Gutmann, and Thomas Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. *Advances in neural information processing systems*, 34:25785–25798, 2021.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Sanaz Kavianpour, James Sutherland, Esma Mansouri-Benssassi, Natalie Coull, and Emily Jefferson. Next-generation capabilities in trusted research environments: interview study. *Journal of Medical Internet Research*, 24(9):e33720, 2022.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andreas Kirsch and Yarin Gal. Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities. *Transactions on Machine Learning Research*.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Steven Kleinegesse and Michael U Gutmann. Bayesian experimental design for implicit models by mutual information neural estimation. In *International conference on machine learning*, pp. 5316–5326. PMLR, 2020.

Kasia Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with llms. In *The Thirteenth International Conference on Learning Representations*.

Jannik Kossen, Cătălina Cangea, Eszter Vértes, Andrew Jaegle, Viorica Patraucean, Ira Ktena, Nenad Tomasev, and Danielle Belgrave. Active acquisition for multimodal temporal data: A challenging decision-making task. *Transactions on Machine Learning Research*.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *International conference on machine learning*, pp. 6450–6459. PMLR, 2021.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Luckeciano C Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. Deep bayesian active learning for preference modeling in large language models. *Advances in Neural Information Processing Systems*, 37:118052–118085, 2024.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela Van Der Schaar, and Pietro Lio. Stochastic encodings for active feature acquisition. In *International Conference on Machine Learning*, pp. 46784–46814. PMLR, 2025.

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

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pp. 4267–4276. PMLR, 2018.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pp. 456–473. PMLR, 2022.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Cohort-based active modality acquisition. *arXiv preprint arXiv:2505.16791*, 2025.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*, 2024.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles. Active learning literature survey. 2009.

Yunyi Shen, Hao Sun, and Jean-François Ton. Reviving the classics: Active reward modeling in large language model alignment. *arXiv preprint arXiv:2502.04354*, 2025.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5972–5981, 2019.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387, 2024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y 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)*, pp. 1500–1519, 2020.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International conference on artificial intelligence and statistics*, pp. 7331–7348. PMLR, 2023.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Yiping Wang, Yifang Chen, Wendan Yan, Alex Fang, Wenjing Zhou, Kevin Jamieson, and Simon S Du. Cliploss and norm-based data selection methods for multimodal contrastive learning. *Advances in Neural Information Processing Systems*, 37:15028–15069, 2024.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54104–54132, 2024.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 155–165, 2025.

Shipeng Yu, Balaji Krishnapuram, Romer Rosales, and R Bharat Rao. Active sensing. In *Artificial Intelligence and Statistics*, pp. 639–646. PMLR, 2009.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A DATA STATISTICS

Here, we report statistics of the pre-training (Figure 4), validation (Figure 5), and test (Figure 7) sets, as well as the pool of radiology reports (Figure 6) used in the experiments of Section 6 of the main text.

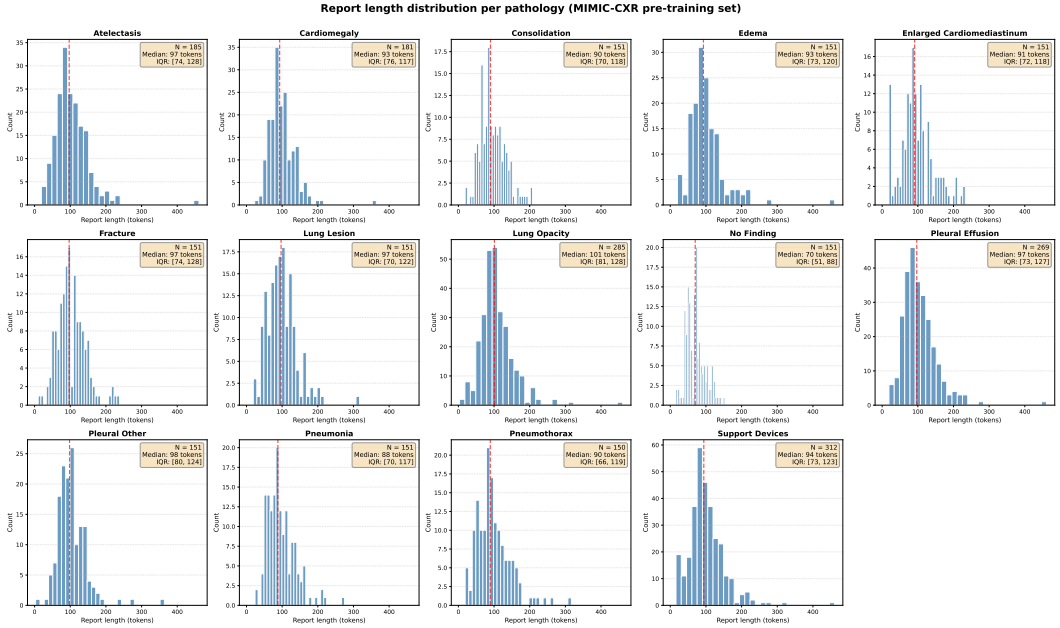


Figure 4: MIMIC-CXR pre-training set ($\mathcal{D}_{\text{pre}}$) statistics.

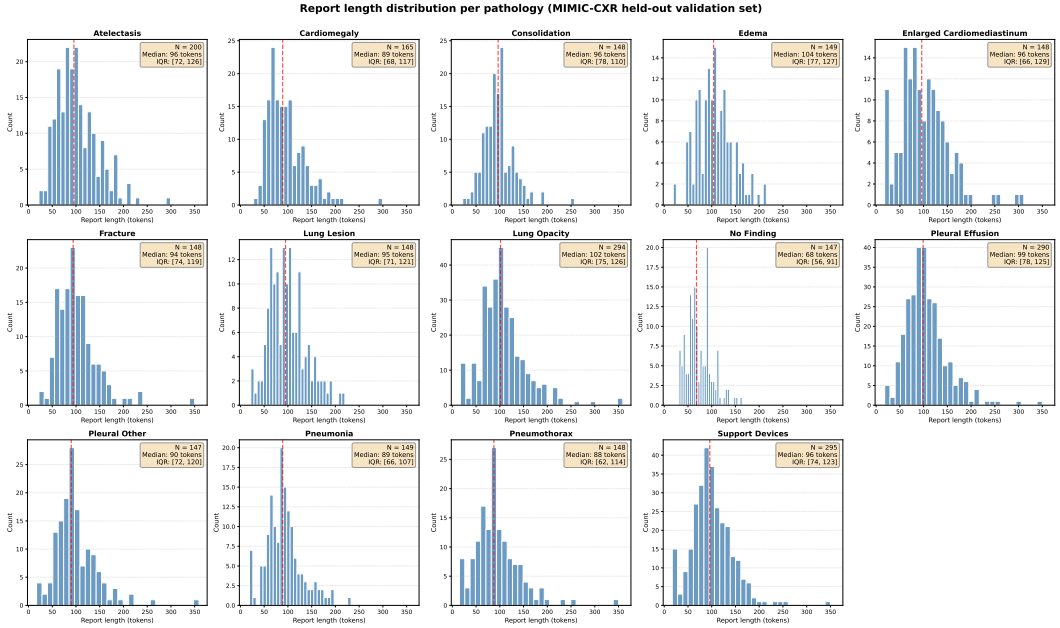


Figure 5: MIMIC-CXR validation set ($\mathcal{D}^*$) statistics.

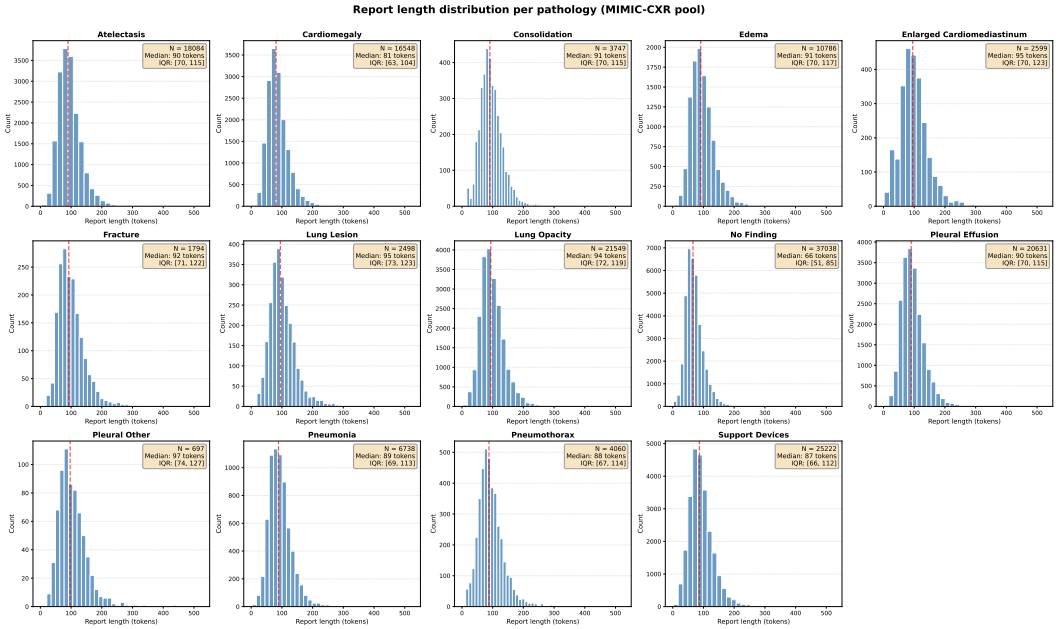


Figure 6: MIMIC-CXR pool ($\mathcal{Y}_{\text{cand}}$) statistics.

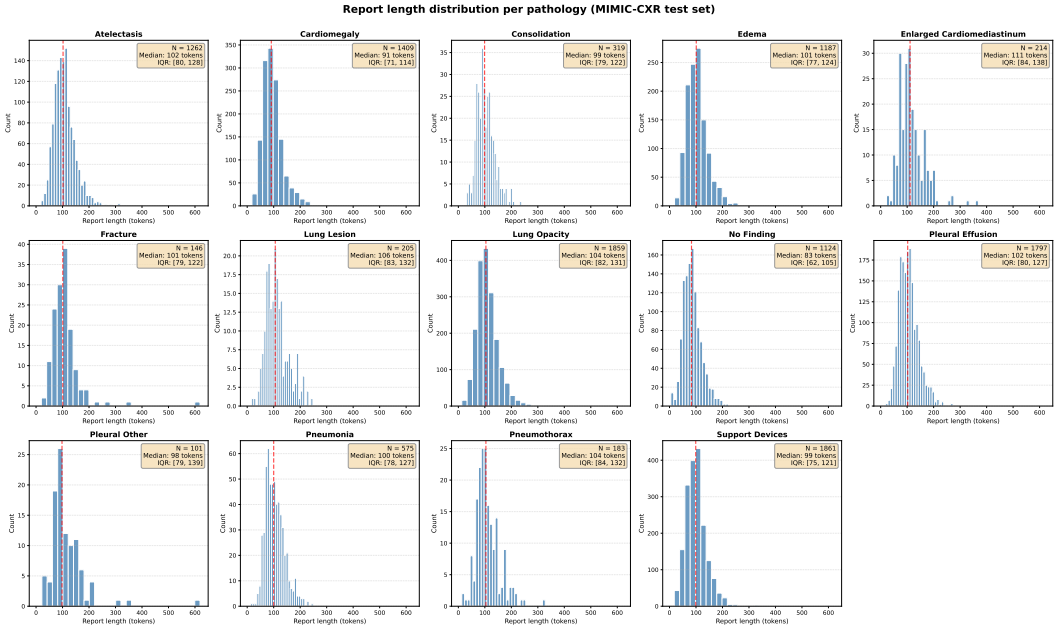


Figure 7: MIMIC-CXR test set ($\mathcal{D}_{\text{test}}$) statistics.

B TRAINING DETAILS

In Table 5, we summarize the hyper-parameters used in our fine-tuning experiments. We use the VLM architecture and the training recipe introduced in (Zambrano Chaves et al., 2025). More specifically, we first optimize the multi-modal projector (Stage 2), and then jointly fine-tune the projector and LoRA layers (Stage 3). Each experiment (i.e., both for data selection and for VLM fine-tuning) was run on 2 NVIDIA H100 80GB GPUs and 16 CPU cores. For report generation, we use deterministic decoding with LLaVA-Rad.

Table 5: VLM training hyper-parameters across different stages. Stage 2 trains only the multi-modal projector (mm_projector), while Stage 3 jointly fine-tunes the projector and LoRA adapters.

Hyper-parameter	**Stage 2**	**Stage 3**
Vision encoder	BiomedCLIP (Zhang et al., 2023) pre-trained on PMC-15M	
Trainable modules	mm_projector (2-layer MLP)	mm_projector + LoRA
Per-device batch size	16	16
Learning rate	1×10^{-3}	1×10^{-4}
LR schedule	cosine decay	cosine decay
Warmup ratio	0.03	0.03
Optimizer	AdamW	AdamW
Precision	bfloat16	bfloat16
Gradient accumulation steps	2	1
Epochs	1	3
LoRA rank	–	64
LoRA alpha	–	128

C ON ACQUISITION FUNCTIONS

In this section, we provide implementation details for each acquisition function defined in Section 4 of the main text.

C.1 GRADIENT-BASED INFLUENCE ESTIMATION

Based on the implementation of Xia et al. (2024), we project gradients onto $d = 8,192$ dimensions using random projections. Furthermore, since Xia et al. (2024) partition the validation set $\mathcal{D}^*$ into subtasks and defines the influence of a candidate as the maximum across subtask-specific scores, we adapt this idea to our setting by partitioning $\mathcal{D}^*$ according to pathology labels $v \in \{1, \dots, V\}$. Following from Equation 4, we get:

$$\begin{aligned}\alpha_{\texttt{LESS}}(\boldsymbol{y}_{\text{cand}}) &= \max_{v \in \{1,\dots,V\}} \text{Inf}_{\text{Adam}}\Big(\boldsymbol{y}_{\text{cand}}, \mathcal{D}^*_{(v)}\Big) \\ &= \max_{v \in \{1,\dots,V\}} \sum_{i=1}^{n} \eta_i \frac{\Big\langle \bar{\nabla}\mathcal{L}\Big(\mathcal{D}^*_{(v)}; \boldsymbol{\theta}_i\Big), \hat{\Gamma}(\tilde{\mathbf{X}}, \boldsymbol{y}_{\text{cand}}; \boldsymbol{\theta}_i)\Big\rangle}{\Big\|\bar{\nabla}\mathcal{L}\Big(\mathcal{D}^*_{(v)}; \boldsymbol{\theta}_i\Big)\Big\| \; \Big\|\hat{\boldsymbol{\Gamma}}(\tilde{\mathbf{X}}, \boldsymbol{y}_{\text{cand}}; \boldsymbol{\theta}_i)\Big\|}.\end{aligned} \tag{8}$$

Analysis of the retrieval-based surrogate. To further examine the quality of the surrogate presented in Formulation 1, here we analyze the retrieval quality. Since the pre-training set contains samples from every pathology, the retrieval process always returns an image sharing at least one pathology label with the candidate report, even though exact matches might not always be available. Quantitatively, the retrieval process achieves a mean Jaccard overlap of 0.875, while 66% of candidate reports were assigned an image with an identical label set. Moreover, even the lowest-overlap retrievals have a Jaccard score of 0.5. We also inspected these lowest-overlap cases and found that the retrieved images always shared at least one pathology label with the candidate reports. The observed mismatches were due to additional co-occurring findings, as those pathology combinations were not present in the pre-training set.

Retrieval-based surrogate for cross-institution acquisition. While Formulation 1 defines the retrieval-based surrogate as a function that retrieves an image from $\mathcal{D}_{\text{pre}}$ based on label overlap with the candidate report $\boldsymbol{y}_{\text{cand}}$, this formulation may not always be practical. For instance, in the cross-institution setting, the label sets of MIMIC-CXR and IU-Xray do not correspond one-to-one. To address this, we propose an alternative surrogate that uses a CLIP text encoder ϕ_{txt} to retrieve the most similar text report from $\mathcal{D}_{\text{pre}}$ to the candidate report $\boldsymbol{y}_{\text{cand}}$ and, in turn, fetch its corresponding image to form the approximate $(\tilde{\mathbf{X}}, \boldsymbol{y}_{\text{cand}})$ pair, i.e.,

$$p_\varsigma(\mathbf{X} \mid \boldsymbol{y}_{\text{cand}}) = \delta(\mathbf{X} - \mathbf{X}_{j^*}), \quad j^* = \underset{j \in \{1,\ldots,|\mathcal{D}_{\text{pre}}|\}}{\arg\max} \phi_{\text{txt}}(\boldsymbol{y}_{\text{cand}})^\top \phi_{\text{txt}}(\boldsymbol{y}_j), \tag{9}$$

C.2 LLM-BASED SCORING

The instruction template for the `Ask-LLM` method is presented in Figure 8, while in Figure 9 we provide the instruction template for its upper-bound variant (`Ask-VLM`).

Ask-LLM

```
You are a vision-language model trained to generate chest X-ray
radiology reports.  You are now selecting additional reports to
expand YOUR OWN ability to generate detailed, accurate, and
clinically precise radiology findings.

Report to evaluate:
###
{report}
###

Would training on this report expand your ability to generate
detailed, accurate chest X-ray findings?

OPTIONS:
- yes
- no
```

Figure 8: The `Ask-LLM` instruction template. The {`report`} field is a placeholder for a candidate radiology report.

C.3 SURROGATE-BASED FEATURE COVERAGE

Here, we define the upper-bound, i.e., image-based, version of the `CLIP` method defined in Section 4 of the main text. More specifically, given an image $\mathbf{X}_{\text{cand}}$ from the pool, the images from the pre-training set $\mathbf{X}_{\text{pre}} \in \mathcal{D}_{\text{pre}}$, and the CLIP vision encoder $\phi_{\text{img}}(\cdot)$, we get the acquisition function `CLIP(img)` defined as follows:

$$\alpha_{\texttt{CLIP}}(\mathbf{X}_{\text{cand}}) = 1 - \max_{\mathbf{X}_{\text{pre}} \in \mathcal{D}_{\text{pre}}} \left(\frac{\phi_{\text{img}}(\mathbf{X}_{\text{cand}})^\top \phi_{\text{img}}(\mathbf{X}_{\text{pre}})}{\|\phi_{\text{img}}(\mathbf{X}_{\text{cand}})\| \|\phi_{\text{img}}(\mathbf{X}_{\text{pre}})\|} \right). \tag{10}$$

C.4 COMPUTATIONAL COMPLEXITY

We analyze the computational complexity of each acquisition function from two angles: the computational burden required to compute intermediate features for each candidate sample, e.g., gradients for the method `LESS` and embeddings for `CLIP` and `FeatEnt`, and the additional overhead arising from the final score computation. Furthermore, to account for the additional cost incurred by `LESS` from the image processing part, we include d_{image} to represent the impact of the vision encoder. In the following, d_{project}, d_{LLM} and d_{CLIP} denote the feature dimensions within which the matching operations are performed, n_{check} denotes the number of VLM checkpoints required by `LESS`, b is the

Ask-VLM (Upper-Bound)

```
<image>

You are a vision-language model trained to generate chest X-ray
radiology reports.  You are now selecting additional data to expand
YOUR OWN ability to generate detailed, accurate, and clinically
precise radiology findings.

You will be given an image and a corresponding report.

Report to evaluate:
###
{report}
###

Would training on this image-report expand your ability to generate
detailed, accurate chest X-ray findings?

Consider both the visual content of the image and the report when
deciding.

OPTIONS:
- yes
- no
```

Figure 9: The `Ask-VLM (Upper-Bound)` instruction template. The `<image>` and {`report`} fields are placeholders for a candidate image and radiology report, respectively.

acquisition batch size, $|\mathcal{Y}_{\text{cand}}|$ is the size of the pool, $|\mathcal{D}^*|$ is the size of the validation set, and $|\mathcal{D}_{\text{pre}}|$ the size of the pre-training set. In Table 6, we observe that `Ask-LLM` is the most computationally efficient method. In practice, it requires only a single forward pass through the LLM, thus bypassing both the costly overhead of the vision encoder and the need for matching operations during score computation. In contrast, `FeatEnt` has the highest computational complexity due to its quadratic dependence on the pool size. However, `LESS` can be more time-consuming in practice due to the cost of processing high-resolution images, computing gradients per candidate sample, and the overhead from averaging multiple model checkpoints.

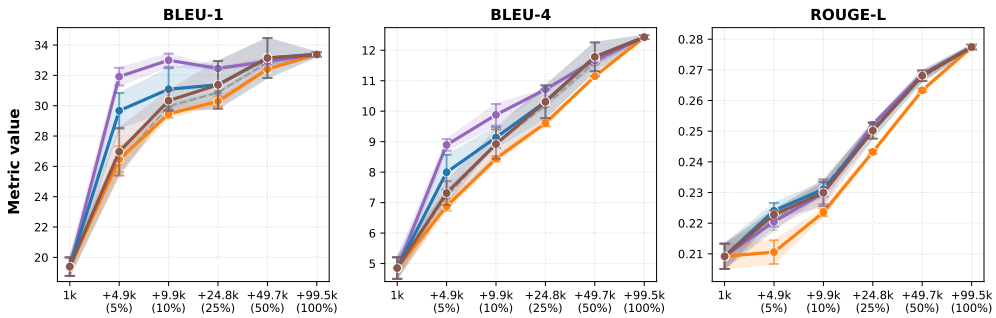
Table 6: Computational complexity of each acquisition method.

Method	Requires Image	Intermediate Feature Computation	Score Computation
`LESS`	✓	$\mathcal{O}(\lvert\mathcal{Y}_{\text{cand}}\rvert \times n_{\text{check}} \times d_{\text{image}})$	$\mathcal{O}(\lvert\mathcal{Y}_{\text{cand}}\rvert \times \lvert\mathcal{D}^*\rvert \times d_{\text{project}})$
`Ask-LLM`	-	-	$\mathcal{O}(\lvert\mathcal{Y}_{\text{cand}}\rvert)$
`CLIP`	-	$\mathcal{O}(\lvert\mathcal{Y}_{\text{cand}}\rvert)$	$\mathcal{O}(\lvert\mathcal{Y}_{\text{cand}}\rvert \times \lvert\mathcal{D}_{pre}\rvert \times d_{\text{CLIP}})$
`FeatEnt`	-	$\mathcal{O}(\lvert\mathcal{Y}_{\text{cand}}\rvert)$	$\mathcal{O}(d_{\text{LLM}} \times \lvert\mathcal{Y}_{\text{cand}}\rvert \times (\lvert\mathcal{Y}_{\text{cand}}\rvert + \lvert\mathcal{D}_{pre}\rvert + b))$

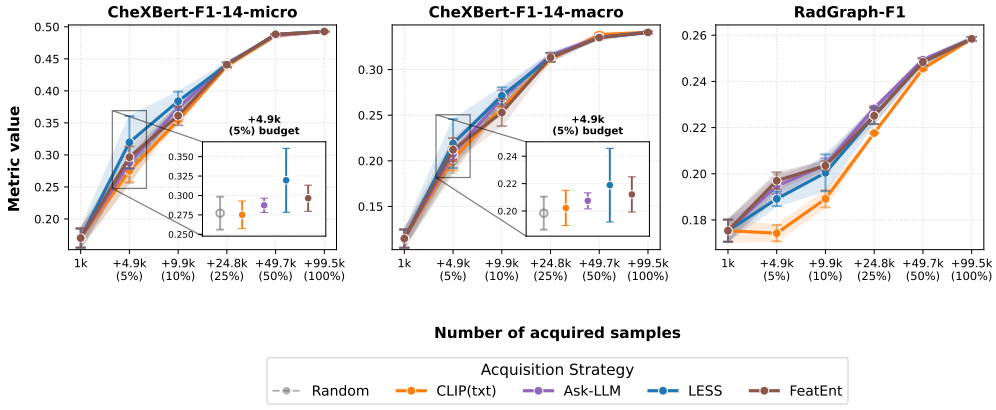
D Additional Results

Extending the analysis presented in Section 6, this section reports complete results across all lexical and clinical metrics for radiology report generation. Figure 10 shows the results of the acquisition experiment on MIMIC-CXR, while the corresponding results on CheXpert-Plus are presented in Figure 11.

On MIMIC-CXR, both `Ask-LLM` and `LESS` yield notable gains over `Random` in BLEU-1–4 (Figure 10a). These differences are less pronounced on CheXpert-Plus (Figure 11a). Moreover, on both datasets, all acquisition methods achieve comparable ROUGE-L and RadGraph-F1 scores.



(a) Lexical evaluation metrics across different acquisition strategies and data budgets.



(b) Clinical evaluation metrics across different acquisition strategies and data budgets.

Figure 10: Stratified acquisition results on MIMIC-CXR. Results are aggregated over 3 random seeds.

Notably, on both datasets, a VLM trained on only 50% of the training data achieves comparable lexical and clinical performance to one trained on the full dataset, indicating that a large portion of the available training data may be redundant.

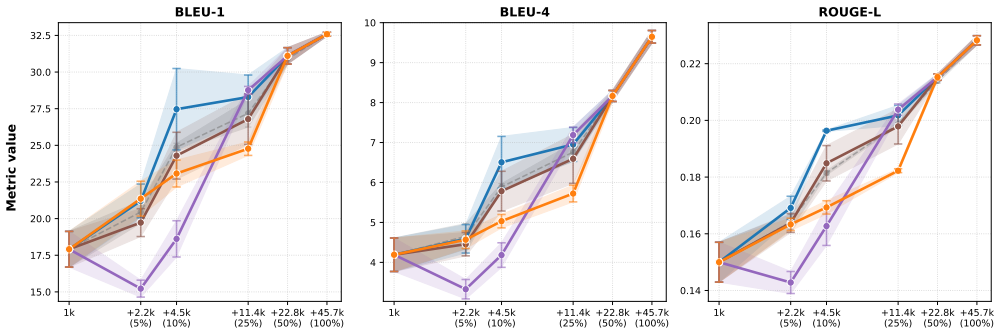
Cross-institution acquisition results. Figure 12 shows CheXbert-F1 scores across multiple data budgets for the two cross-institution acquisition experiments considered in this work. It is worth noting that, even in this challenging setting, the performance of full-data training can be recovered using only a fraction of the available data, e.g., 25%, showing that the remaining data provides limited additional benefit.

E Additional Experiments

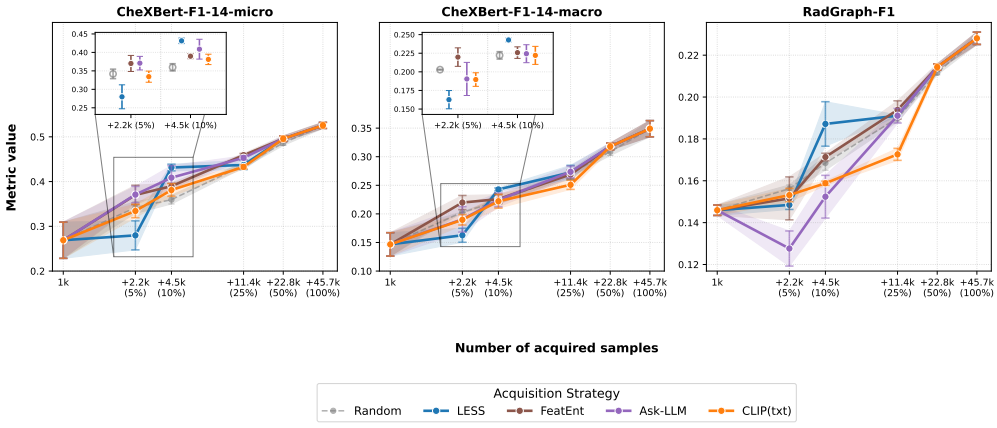
In this section, we provide additional analysis of the results presented in Section 6 of the main text.

E.1 Diversity of selected images

To further analyze the behavior of each acquisition method, we examine the diversity of the images corresponding to the selected text reports in terms of the average pairwise cosine distance of their CLIP image embeddings, measured both within each selected subset and with respect to the pre-training set images. As shown in Figure 13, and along with the findings presented in Figure 3, we find that while the `CLIP(img)` method (i.e., the upper-bound version of our `CLIP(txt)` method) yields substantially more diverse images than the others, this does not translate into higher downstream performance. Moreover, the remaining methods—including random acquisition—select comparably



(a) Lexical evaluation metrics across different acquisition strategies and data budgets.



(b) Clinical evaluation metrics across different acquisition strategies and data budgets.

Figure 11: Stratified acquisition results on CheXpert-Plus. Results are aggregated over 3 random seeds.

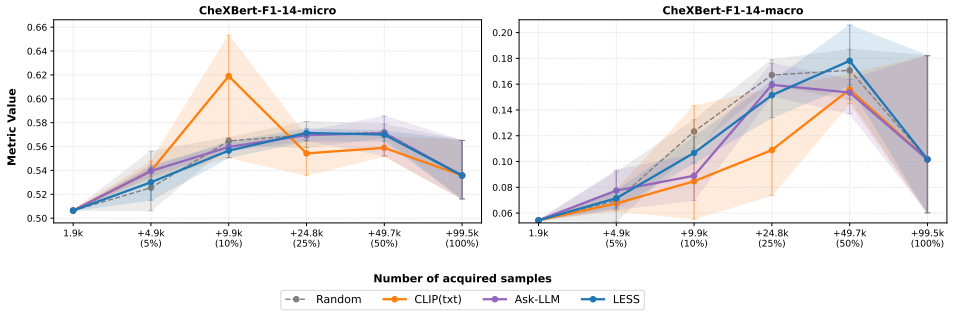
diverse images. Overall, these findings suggest that image diversity in the CLIP embedding space alone is not an effective data selection criterion for the downstream task of report generation.

E.2 Overlap between acquisition methods

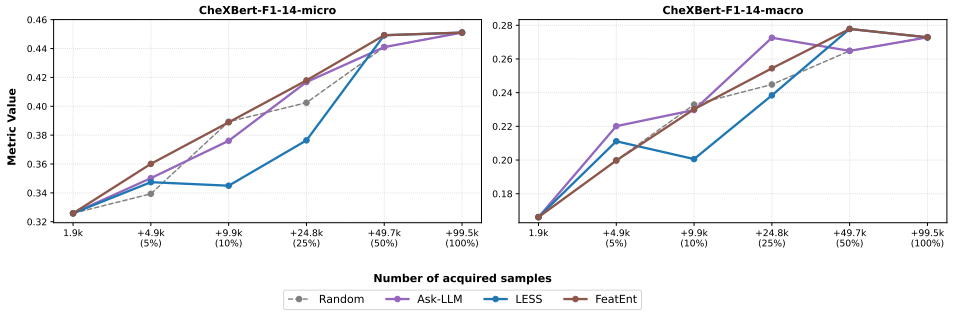
We analyze the overlap between samples selected by different acquisition methods to assess redundancy across strategies. As shown in Figure 14, the selected subsets exhibit moderate overlap ($\sim$ 30%) at a 5% data budget, indicating that the methods already agree on a subset of highly informative samples while also capturing complementary information. As the budget increases to 10%, the pairwise overlap reaches approximately 45% on average, suggesting that the methods start to converge towards a shared set of high-utility text reports even at relatively small data budgets.

E.3 Report length bias analysis

We further examine whether the acquisition methods exhibit a systematic bias with respect to report length. As shown in Figure 15, the `Ask-LLM` method consistently assigns higher acquisition scores to more verbose reports across pathologies. In contrast, `LESS` shows the opposite trend, often assigning higher scores to shorter reports within certain pathologies, particularly in the case of *No Finding*.



(a) Resuts for the IU-Xray/MIMIC-CXR setting. We report CheXbert-F1 scores on the IU-Xray test set. Results are aggregated over 3 random seeds.



(b) Results for the CheXpert-Plus/MIMIC-CXR setting. We report CheXbert-F1 scores on the CheXpert-Plus test set.

Figure 12: Cross-institution acquisition results.

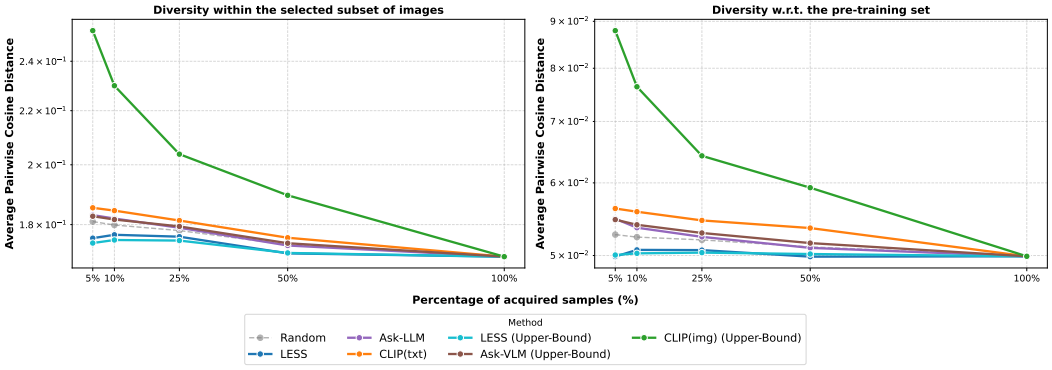


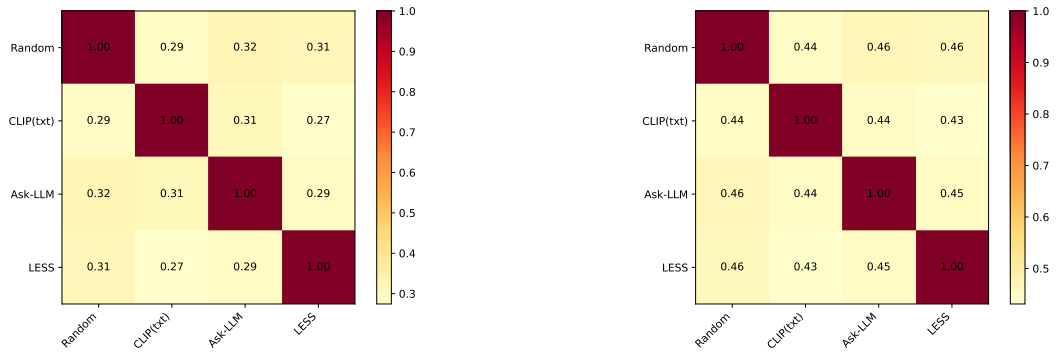
Figure 13: Diversity of selected images.

F Ablation Studies

In this section, we present ablations of key components of our text-based active image selection pipeline, including the size of the pre-training set $|\mathcal{D}_{\text{pre}}|$ and the choice of acquisition strategy, i.e., naive top-k vs. stratified top-k.

F.1 Effect of the pre-training set size $|\mathcal{D}_{\text{PRE}}|$

To assess the sensitivity of acquisition methods to the amount of available pre-training data, we repeat the acquisition experiments on MIMIC-CXR (Section 6 of the main text) with $|\mathcal{D}_{\text{pre}}| = 10,000$. The results in Figure 16 show that all methods achieve comparable clinical performance, with only



(a) 5% Budget (b) 10% Budget

Figure 14: Comparison of overlap between acquisition methods at different data budgets.

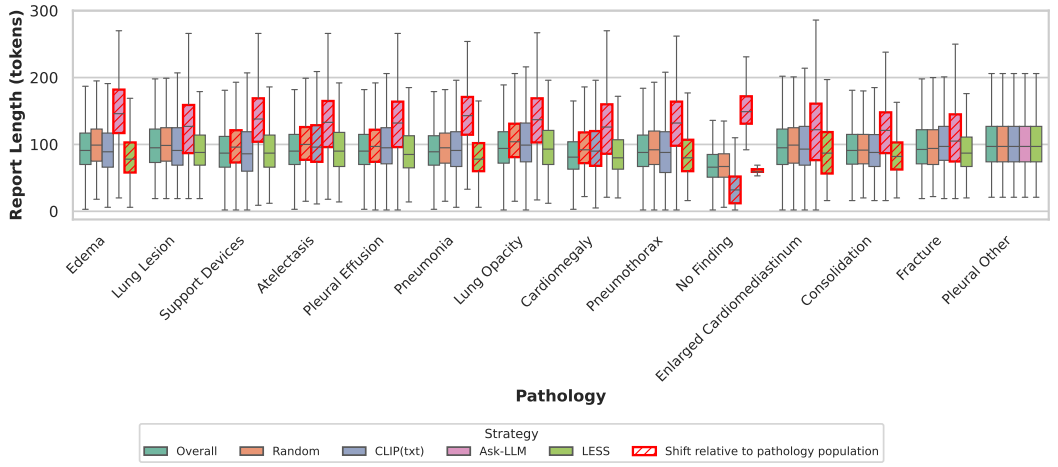
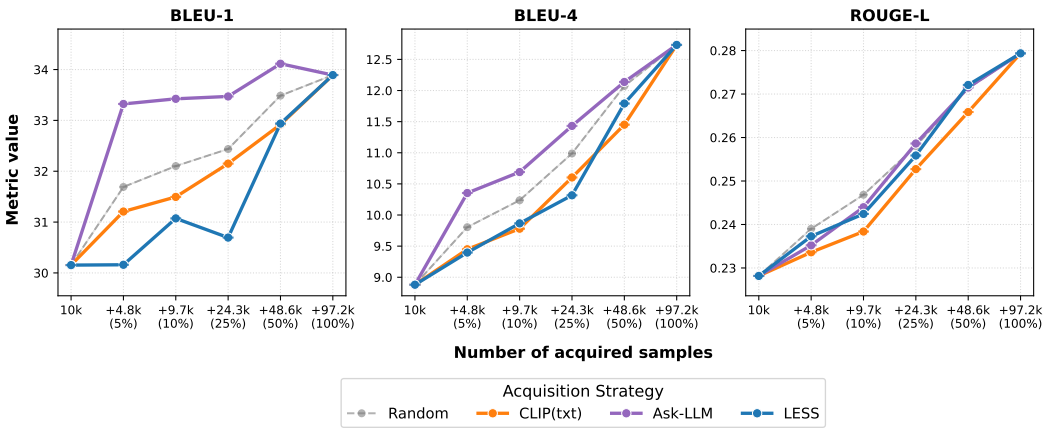


Figure 15: Assessment of report length bias for a 5% data budget. For each method and pathology subgroup, we compute Cliff's δ (Cliff, 1993) between the report lengths of selected samples and the full report length distribution of that pathology subgroup, where a positive δ indicates a tendency toward selecting longer reports. Subgroups with $|\delta| > 0.10$ are flagged as exhibiting a practically significant length bias.

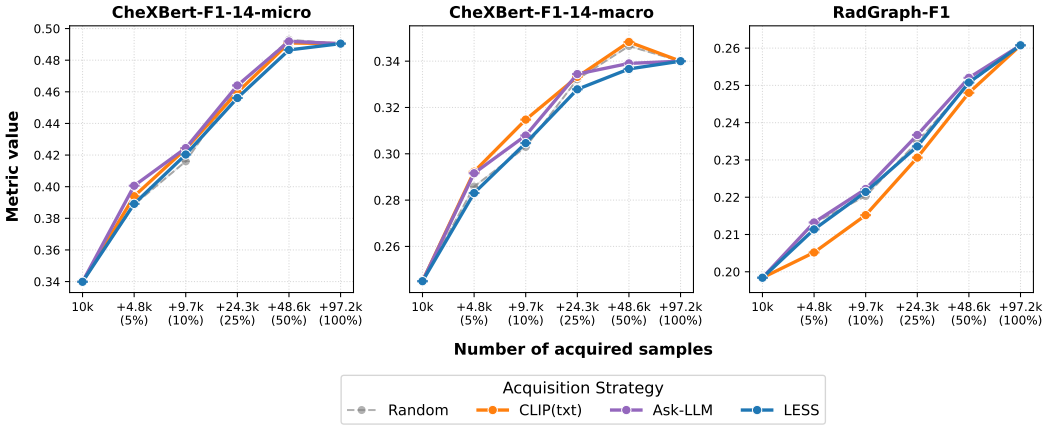
small differences relative to random acquisition across budgets. This suggests that, in this regime, performance may be less sensitive to the specific acquisition strategy, and that data selection yields limited gains when sufficient pre-training data is available.

F.2 Naive vs. Stratified top-k Acquisition

In Section 6 of the main text, we report results using a stratified acquisition strategy to account for label imbalance. Given that MIMIC-CXR data is multi-label, we implement stratified acquisition as an iterative greedy sampling process, where at each step we select the highest-scoring sample per pathology according to the acquisition score assigned by each method. To motivate this design choice, we evaluate how much each acquisition method benefits from stratification compared to a naive top-k approach that does not account for label information. As shown in Figure 17, stratification consistently improves the performance of acquisition methods in terms of CheXBert-F1 score (macro averaged), especially in the case of rare pathologies in the test set.



(a) Lexical report generation evaluation metrics across acquisition strategies and data budgets.



(b) Clinical report generation evaluation metrics across acquisition strategies and data budgets.

Figure 16: Stratified acquisition results on MIMIC-CXR with $|\mathcal{D}_{\text{pre}}| = 10,000$.

G LICENSES FOR EXISTING ASSETS

Here, we provide the licenses for existing assets used in our work:

- The implementation of the LLaVA-Rad (Zambrano Chaves et al., 2025) training framework is available in `https://github.com/microsoft/LLaVA-Rad` (under Microsoft Research License).
- The core implementation of LESS (Xia et al., 2024) is available in `https://github.com/princeton-nlp/LESS` (under MIT License).
- The core implementation of `FeatEnt` (Melo et al., 2024) is available in `https://github.com/luckeciano/BAL-PM`.
- The MIMIC-CXR dataset, as well as LLaVA-Rad MIMIC-CXR annotations, are available with credentialed access through PhysioNet `https://physionet.org/` (under PhysioNet Credentialed Health Data License 1.5.0).

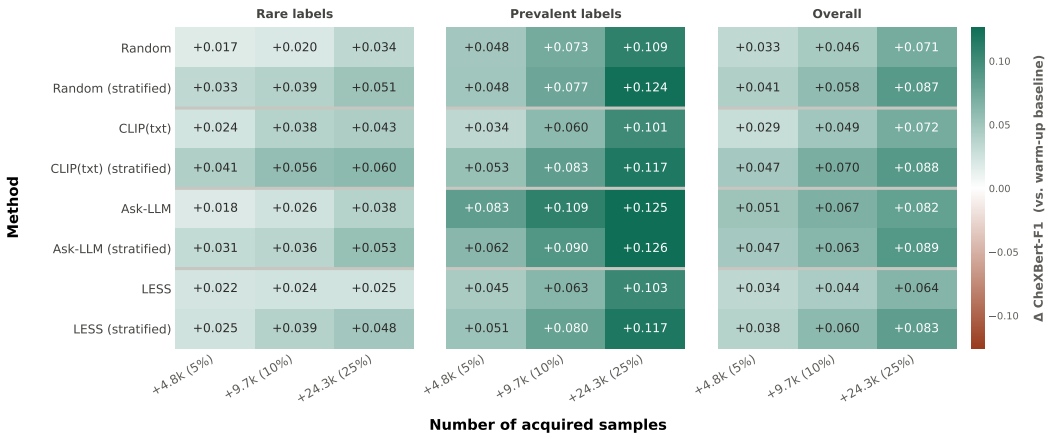


Figure 17: Comparison between naive vs. stratified top-k acquisition across budgets. Note that the warm-up baseline is the one presented in Appendix F.1 with $|\mathcal{D}_{\text{pre}}| = 10,000$.