SemEnrich: Self-Supervised Semantic Enrichment of Radiology Reports for Vision-Language Learning

Halil Ibrahim Gulluk
Department of Electrical Engineering
Stanford University
gulluk@stanford.edu

Olivier Gevaert
Department of Biomedical Data Science
Stanford University
ogevaert@stanford.edu

Abstract

Medical vision–language datasets are limited in size and biased toward abnormal findings: clinicians document abnormalities in detail but frequently omit normal observations because they are considered uninformative for the patient's condition. We propose a self-supervised enrichment method that leverages semantic clustering of report sentences and augments the training reports with normal observations drawn from co-occurring clusters in the same dataset. Our approach yields consistent gains in supervised fine-tuning: average improvements of $+5.63\%$ COMET, $+3.04\%$ BERTScore, $+7.40\%$ Sentence-BLEU, $+5.30\%$ CheXbert-F1, and $+7.47\%$ RadGraph-F1 across five language model backbones. Ablation studies confirm that improvements stem from semantic clustering rather than random augmentation. We further introduce a cluster-aware reward for GRPO training that complements exact-match rewards and yields additional gains: $+2.78\%$ COMET, $+3.14\%$ BERTScore, and $+14.21\%$ ROUGE-L. Code is available at https://github.com/igulluk/SemEnrich.

1 Introduction

Vision–language models trained on large-scale web data have made rapid progress on image captioning, visual question answering, and text-to-image generation (Radford et al., 2021; Li et al., 2023b; Jia et al., 2021). The dominant lesson from this line of work is simple: the larger and richer the training data, the better the downstream performance.

The medical domain, however, breaks this scaling assumption. Public medical image–text corpora are small, expensive to curate, and biased in ways general-domain data is not. Many medical visual question answering benchmarks contain only coarse prompts (*e.g.*, "which organ is presented" or "which imaging modality is shown"), which are insufficient to evaluate models intended for real clinical use (Lau et al., 2018; Zhang et al., 2024). Building new medical datasets is one response; improving existing ones is the complementary, lower-cost path that we pursue in this work.

To make this concrete, we focus on chest X-ray reports, which contain two main sections: *findings* and *impressions*. The findings section is a list of observations made by the radiologist from the image; the impression is the radiologist's final conclusion, often informed by patient history and clinical indication in addition to the image itself. Because the impression depends on information outside the image, we target the findings section, where a large vision–language model (LVLM) can in principle ground every statement in the visual input.

The findings section also has a structural property that ordinary text data does not: its sentences are largely independent. A radiologist may write the same set of observations in any order without changing the meaning of the report. This order invariance lets us treat the findings as an *unordered set* of observations and, after embedding each sentence, group them into semantic clusters and reason about which clusters can co-occur.

A second property of these reports motivates our enrichment. When writing a report, clinicians focus on abnormalities and tend to omit normal observations that they consider uninformative for the patient's condition, so the training data systematically under-represents normal anatomy. Crucially, the two properties above let us repair this bias without external supervision: because findings are an exchangeable set, a normal observation that has been omitted can be reinserted as long as it is consistent with what the report already states.

Our approach. We introduce **SemEnrich**, a self-supervised pipeline that exploits both properties. We embed every finding sentence in the training corpus with a sentence transformer and group the embeddings into semantic clusters. Each

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39416–39443
October 24-29, 2026 ©2026 Association for Computational Linguistics

cluster is assigned a *polarity* (normal or abnormal), and two clusters are deemed *connected* when they already co-occur in some real report. For a given patient we then add only normal clusters that are mutually connected to every cluster already present – a *valid enrichment* – so that the inserted sentences are plausibly omitted normal findings rather than fabricated content. The same clustering is reused as a reward signal for GRPO post-training, where a cluster-level F1 reward complements the standard exact-match reward. We apply this type of self-supervision to the training set and the test set is never modified, so all reported gains reflect a richer training signal rather than easier targets.

2 Related Work

Medical vision–language models: A growing line of work targets medical visual question answering and radiology report generation under limited data. Lin et al. (2023) curate a medical image–caption dataset and train a CLIP-style model; Wu et al. (2024) train a Llama model for medical multiple-choice question answering. LLaVA-Med (Li et al., 2023a) extends LLaVA to medical conversations including images, and R-LLaVA (Chen et al., 2025) improves medical VQA with region-of-interest grounding. Tu et al. (2024) propose a generalist medical model that handles VQA, report generation, and classification within a single architecture.

Improving information extraction from medical data: Several approaches focus on richer alignment or retrieval rather than larger data. MMed-RAG (Xia et al., 2024) introduces a domain-aware multi-modal RAG pipeline with preference fine-tuning. Med-UniC (Wan et al., 2023) aligns medical text across English and Spanish to broaden coverage; Boecking et al. (2022) show that combining global and local alignment of visual and text representations improves downstream medical VLM performance.

Datasets and report generation models: We use ReXGradient-160K (Zhang et al., 2025), which contains 160K chest X-ray studies across 109,487 patients from 79 medical sites; the rationale for this choice is discussed in Appendix A.15. Related releases include 3D-RAD (Gai et al., 2026) for 3D medical VQA and RadVLM (Deperrois et al., 2025) for conversational radiology. For report generation specifically, Flamingo-CXR (Tanno et al., 2025) adapts a Flamingo backbone with expert evaluation in the loop, Chen et al. (2020) propose a memory-driven transformer for radiology report generation, and Alfarghaly et al. (2021) use conditioned transformers for automated report generation.

In contrast to these methods, SemEnrich does not propose a new dataset or a new architecture. It produces an *enrichment* of any existing chest X-ray dataset, by adding semantically consistent normal findings that the radiologist did not bother to mention, and it can be combined with any of the report-generation backbones above.

3 Method

This section describes (i) how we cluster the sentences in the findings section across all training reports, (ii) how we label each cluster as normal or abnormal, and (iii) how we use this structure to enrich the training findings in a self-supervised manner.

3.1 Sentence Representations and Clustering

We operate on the findings section of each report, which lists observations made directly from the chest X-ray.

We embed every finding sentence in the training set with a pretrained sentence transformer, `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019), denoted f_{sent}. Let N be the total number of unique finding sentences across all training reports. The encoder produces N representations $\tilde{x}_i = f_{sent}(s_i) \in \mathbb{R}^{384}$ for $i = 1, \ldots, N$, which we normalize to unit length, $x_i = \tilde{x}_i / \|\tilde{x}_i\|$.

For the report in Figure 2, the findings decompose into seven sentences $s_1, \ldots, s_7$: for instance, s_1 is "The lungs are adequately inflated", s_2 is "There is no focal infiltrate", and s_7 is "The bony thorax exhibits no acute abnormality". The corresponding embeddings $x_1, \ldots, x_7$ feed into the next step, whose goal is to group sentences so that each group contains semantically equivalent observations.

Clustering finding sentences: With these embeddings in hand, we cluster the sentences so that semantically equivalent observations fall into a common group. We use three algorithms, all with Euclidean distance: K-means, DBSCAN, and HDBSCAN (Lloyd, 1982; Ester et al., 1996; Campello et al., 2013). With default hyperparameters, DBSCAN yields 995 clusters and HDB-

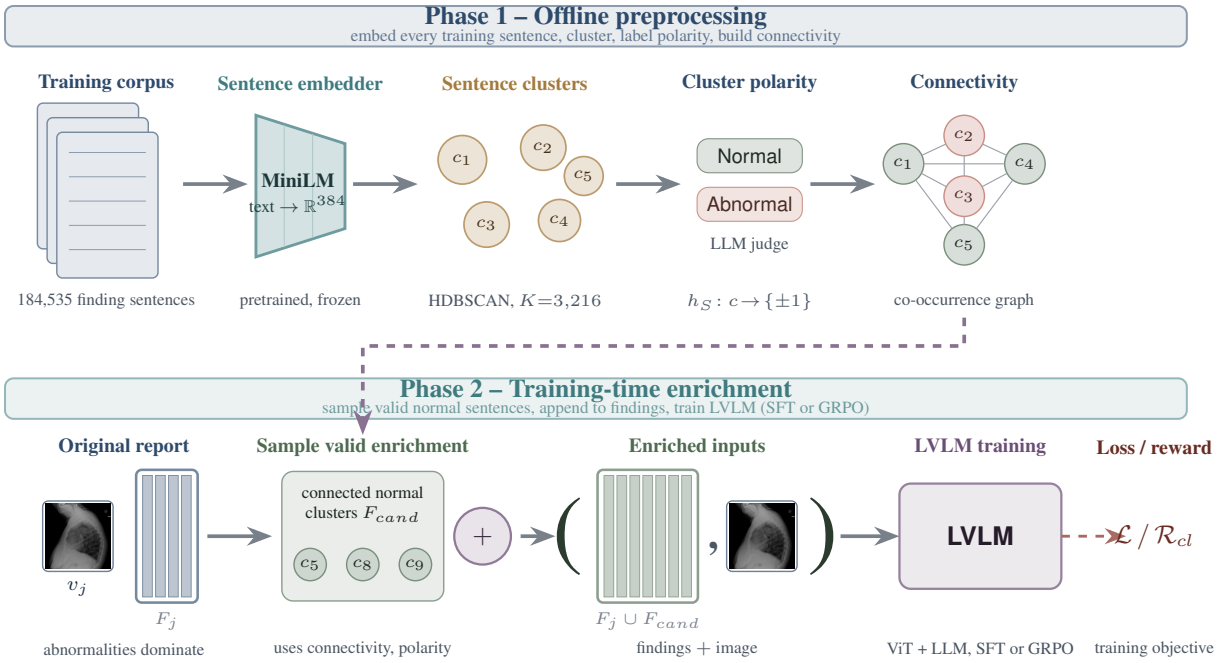


Figure 1: **SemEnrich pipeline.** *Phase 1 (top):* every finding sentence in the training set is embedded by a frozen sentence transformer (MiniLM) and clustered with HDBSCAN. Each cluster is labeled *normal* or *abnormal* by an LLM, and cluster co-occurrence across patients defines a connectivity graph. *Phase 2 (bottom):* for every patient (v_j, F_j) we sample a *valid enrichment*—a set of normal clusters F_{cand} connected to every cluster already in F_j—and append one sentence per added cluster to the findings. The enriched findings and the chest X-ray then train an LVLM via SFT or GRPO; in the GRPO setting the same clustering is reused as a cluster-F1 reward $\mathcal{R}_{cl}$.



Figure 2: An example chest X-ray from the ReXGradient-160K dataset. **Findings:** The lungs are adequately inflated. There is no focal infiltrate. There is no pleural effusion. No pulmonary parenchymal nodules or masses are observed. The heart and pulmonary vascularity are normal. There is calcification in the wall of the thoracic aorta. The bony thorax exhibits no acute abnormality. **Impression:** No pulmonary nodules are observed. There is no acute cardiopulmonary abnormality. Thoracic aortic atherosclerosis.

SCAN yields 3,216; for K-means we sweep $k \in \{1000, 2000, 5000\}$.

The two families behave differently. Density-based methods (DBSCAN, HDBSCAN) leave low-density points as *outliers*, so not every sentence is assigned to a cluster. K-means partitions all points and is forced to absorb outliers into the nearest cluster. In our data, K-means assigns all 184,535 training sentences to clusters, whereas HDBSCAN clusters only 66,477 of them; the remaining sentences are dropped from the enrichment pool. This trade-off matters: density-based methods produce purer clusters at the cost of coverage, while K-means trades purity for coverage. The number of clusters in K-means controls a similar trade-off between expansion reach and cluster noise. To avoid hyperparameter cherry-picking, all clustering methods use library defaults and we report all three values of k for K-means. The choice of sentence encoder is itself a design axis: in Appendix A.9 we hold the pipeline fixed and swap MiniLM for larger general-purpose encoders (Stella, GTE-Qwen2) and biomedical ones (PubMedBERT, BioBERT); every encoder improves over the baseline, and the general-purpose MiniLM remains the strongest on the clinical metrics.

Below we provide examples of sentences that are grouped into the same cluster, demonstrating that semantically similar findings are successfully clustered together:

Cluster 1: Gray-White Differentiation

"Gray-white differentiation appears unremarkable." / "The gray-white matter differentiation is within normal limits." / "Gray white differentiation is intact." / ...

Additional cluster examples are provided in Appendix A.1.

Let $f_C : s_i \mapsto c_i$ denote the sentence-to-cluster-id mapping, where c_i is the cluster id assigned to sentence s_i. With K clusters in total, $C = \{c_1, \ldots, c_K\}$, the findings of the j-th chest X-ray, which is naturally a set of sentences $\{s_{j1}, s_{j2}, \ldots, s_{jr}\}$ where r is the number of sentences in that report, can therefore be represented as the corresponding set of cluster ids $\{c_{j1}, c_{j2}, \ldots, c_{jr}\} = \{f_C(s_{j1}), \ldots, f_C(s_{jr})\}$. Per-method cluster-size statistics are reported in Appendix A.3 an A.4.

3.2 Cluster Polarity Assignment

We assign each cluster a polarity label: *normal* or *abnormal*. A cluster is normal if its sentences describe healthy, unsuspicious anatomy or non-findings (*e.g.*, "the lungs are clear"); it is abnormal if its sentences describe a pathological, suspicious, or symptomatic finding (*e.g.*, "there is a left lower lobe consolidation").

To label a cluster, we sample one sentence from it at random and prompt a DeepSeek-R1 model with the following instruction:

```
Decide whether the following radiology sentence describes a normal observation or an abnormal observation.
- If the sentence reports any pathology, suspicious feature, or symptomatic finding, respond with 'abnormal'.
- If the sentence describes a finding consistent with a healthy patient, respond with 'normal'.
Output only 'normal' or 'abnormal', with no explanation.
The sentence is: {finding}
```

We repeat this for every cluster, producing a polarity map $h_S : c_i \mapsto s_i$ with $s_i \in \{+1, -1\}$, where $+1$ denotes a normal cluster and -1 denotes an abnormal one.

On the noise of single-sentence polarity assignment: This single-sentence labelling is noisy and is not a substitute for clinician judgement. It is tolerable here because polarity acts only as a hard *gate* on which clusters may be added (Section 3.3), never as a supervision target: a wrong label only changes whether a cluster is eligible, and mis-labelled clusters are further filtered by the co-occurrence constraint, so errors lower the augmentation rate rather than inject incorrect content. Empirically, even under this noisy gate the gains in Section 4 are large and consistent, indicating that the signal dominates the residual noise; clinician-in-the-loop labelling is a natural refinement. We also directly quantify this noise by labelling member sentences of every inserted normal cluster: the inserted clusters are highly polarity-pure (insertion-weighted purity $\approx 99.8\%$), corresponding to an effective abnormal-injection rate of only $\approx 0.22\%$ (Appendix A.11).

3.3 Enrichment via Connected Clusters

We now describe how new sentences are added to a patient's findings. Three definitions formalize the procedure.

Connected Clusters: Two clusters c_i and c_j are *connected* if there exists a patient in the training set whose findings contain one sentence from c_i and another from c_j. Connectivity is a co-occurrence relation: if two clusters are connected, then the propositions they represent have already been seen together in at least one real report and are not mutually contradictory in that context. In practice we apply a slightly more robust co-occurrence gate governed by two thresholds, fixed across all configurations: two clusters are connected only when they co-occur in at least $\tau_{\text{count}} = 3$ reports (a floor on the raw co-occurrence count) and their row-normalized co-occurrence exceeds $\tau_{\text{norm}} = 0.001$. This discards spurious single-report co-occurrences; the full procedure is given in Appendix A.5, Algorithm 1.

Valid Enrichment: Let $F_j = \{c_{j1}, c_{j2}, \ldots, c_{jr}\}$ be the set of clusters spanned by the findings of the j-th patient. For a candidate set $F_{cand} \subset C$, we call (F_j, F_{cand}) a *valid enrichment* if:

- $F_j \cap F_{cand} = \emptyset$, (1)
- $\forall\, x, y \in F_j \cup F_{cand},\ x, y$ are connected, (2)
- $\forall\, c \in F_{cand},\ h_S(c) = +1$. (3)

Intuitively, a valid enrichment adds only *normal* observations to an existing report, and only when every added cluster co-occurs with every other cluster already present or being added. The motivation is that clinicians focus on abnormalities and frequently omit normal observations they consider uninformative; by definition we never add abnormalities, and connectivity prevents combining mutually contradictory normal findings.

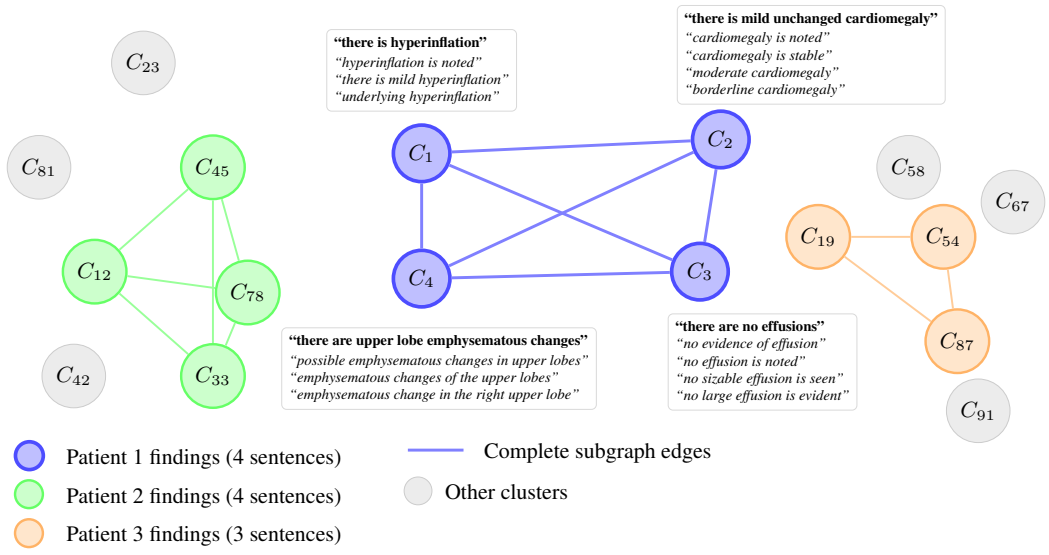


Figure 3: Visualization of the cluster-based findings representation. Each node represents a cluster of semantically similar sentences. Patient 1's findings (consisting of multiple sentences) forms a complete subgraph where each node corresponds to the cluster containing one of the finding sentences. The blue subgraph shows an example findings with 4 sentences: the bold text shows the original sentence, and italicized text shows other sentences from the same cluster. The green and orange subgraphs represent findings from other patients.

Largest Valid Enrichment: (F_j, F_{cand}) is a *largest valid enrichment* if it is a valid enrichment and no additional cluster can be added without breaking the conditions above:

- (F_j, F_{cand}) is a valid enrichment, (4)
- $\forall\, c \in C \setminus (F_j \cup F_{cand})$, $(F_j \cup F_{cand}, \{c\})$ is not a valid enrichment. (5)

In other words, we add as many normal sentences as possible without violating connectivity or polarity. Figure 4 illustrates a small example. Finding possible valid enrichments is very closely related to classical *maximal clique enumeration* problem in graph theory. The algorithmic details of finding the largest valid enrichment are deferred to Appendix A.5.

Like the polarity assignment, this enrichment is inherently noisy; however, the large and consistent improvements in Section 4 experimentally show that the signal it provides far outweighs that noise.

Enrichment examples: We now show a real example of a valid enrichment from our training data, pairing an original set of findings with the expansion clusters that the procedure adds. The inserted sentences are semantically consistent with the original report, and the procedure never introduces an abnormal finding.

Enrichment Example (Multiple Expansions)

Initial Findings:

- Right hilar mass is suspected.
- Heart is normal in size. Lungs are clear.
- No pneumothorax or effusion.
- Airspace disease is present in the inferior right upper lobe.

Expansion Cluster 1: *"visualized bony thorax intact", "the visualized bony structures of the thorax are intact", ...*

Expansion Cluster 2: *"normal lung volumes", "lung volumes are normal", "the lung volumes are normal", ...*

The same initial findings can yield several distinct valid enrichments depending on which connected normal clusters are sampled; further alternatives and examples are given in Appendix A.2. Appendix A.10 additionally reports which normal findings the procedure most often recovers – canonical omissions such as *normal heart size* and *no pneumothorax* – together with the fraction of reports that are enriched.

3.4 Training: SFT and Cluster-Aware GRPO

Supervised fine-tuning: The backbone follows the standard vision–language recipe (Figure 5): a Vision Transformer encodes the chest X-ray and a linear layer projects its features into m image

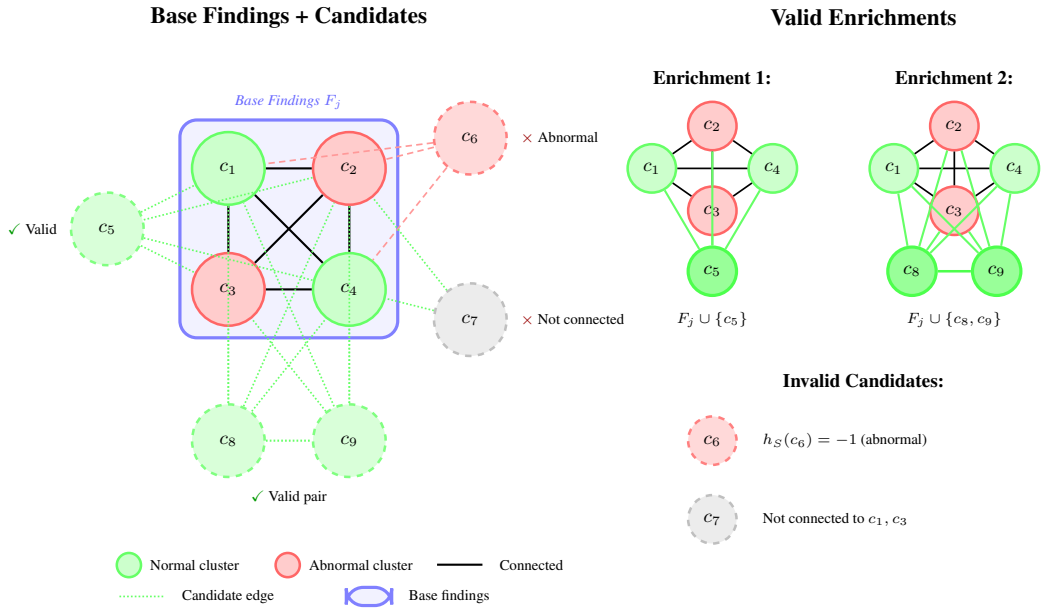


Figure 4: Illustration of the valid enrichment process. **Left:** the base findings $F_j = \{c_1, c_2, c_3, c_4\}$ (inside the blue box) and candidate clusters (dashed nodes). Green nodes are normal clusters; red nodes are abnormal clusters. Candidate c_5 is valid (normal and connected to all base clusters). Candidate c_6 is invalid (abnormal). Candidate c_7 is invalid (not connected to c_1 and c_3). Candidates c_8 and c_9 form a valid pair (both normal, connected to all base clusters and to each other). Note that c_5 is not connected to $\{c_8, c_9\}$, so they cannot belong to the same enrichment. **Right:** two largest valid enrichments, $F_j \cup \{c_5\}$ and $F_j \cup \{c_8, c_9\}$. $F_j \cup \{c_8\}$ is a valid enrichment but not a *largest* one.

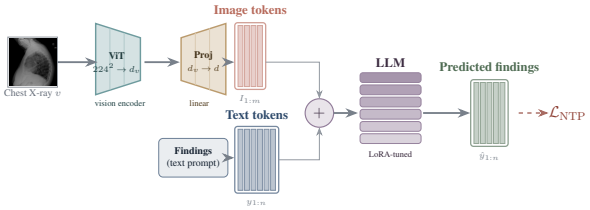


Figure 5: Model architecture. The chest X-ray is encoded by a Vision Transformer and mapped by a linear projector into m image tokens $I_{1:m}$. These are concatenated with the tokenized findings $y_{1:n}$ and passed to a LoRA-tuned large language model, trained with the next-token prediction loss on the finding tokens.

tokens $I_{1:m}$, which are concatenated with the tokenized findings $y_{1:n}$. We train with the standard next-token-prediction objective, masking the image tokens in the loss so the model is only penalized for predicting finding tokens conditioned on the image representation:

$$\mathcal{L} = -\frac{1}{n}\sum_{t=1}^{n} \log P_\theta(y_t \mid I_{1:m},\, y_{1:t-1}), \qquad (6)$$

where θ are the trainable parameters and $P_\theta(\cdot)$ is the LLM's next-token distribution.

Cluster-aware GRPO reward: Our clustering can also be reused as a reward signal in RL post-training, independently of enrichment. We adopt the same framing as GRPO over reasoning models: the input is formatted as `<think>findings</think> <answer>impression</answer>`, and the standard reward is the exact match between the generated and ground-truth impressions, $\mathcal{R}_{exact} = \mathbf{1}[y_{gt} = y_{gen}]$.

We add a complementary reward based on the cluster assignments of the generated findings. Let $F_{gt} = \{c_1, \ldots, c_l\}$ denote the set of clusters spanned by the ground-truth findings of a sample, and $F_{gen} = \{c_{g1}, \ldots, c_{gm}\}$ the set spanned by the generated findings. The cluster-aware reward is the F1 between these two sets, added to the exact-match reward:

$$\mathcal{R}_{cl} = \frac{2\,|F_{gt} \cap F_{gen}|}{|F_{gt}| + |F_{gen}|} + \mathcal{R}_{exact}. \qquad (7)$$

In experiments, we compare $\mathcal{R}_{cl}$ to $\mathcal{R}_{exact}$ alone. This is orthogonal to the enrichment in Section 3.3: enrichment changes the training data, while the cluster-aware reward changes the supervision signal in RL.

4 Experiments

We evaluate SemEnrich on the ReXGradient-160K dataset (Zhang et al., 2025). We use the official patient-disjoint splits, with train/validation/test = 95,716 / 6,964 / 6,807 patients and zero patient

overlap between splits. Clustering, cluster polarities, and the co-occurrence connectivity graph are all built from the *training* split only. For each configuration, a baseline model is trained on the raw dataset, and an enriched model is trained on the same dataset with our augmentation applied at batch time: for every sample, we draw uniformly at random one of its precomputed largest valid enrichments and append the corresponding sentences to the findings before the forward pass. The evaluation splits are never enriched and the test set is never modified, so there is no train–test leakage and all the performance improvements experimentally prove a richer self-supervision.

Backbones: We use ViT-B/16 as the vision encoder with input resolution 224×224. We evaluate five LLM backbones: DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025), Qwen3-8B (Yang et al., 2025), Mistral-7B-v0.1 (Jiang et al., 2023), and Gemma3-4B (Gemma Team, 2025).

Training: We fine-tune with LoRA (rank r=8) for 25 epochs using AdamW with learning rate 10^{-4} and batch size 144. All experiments run on a single NVIDIA A100 GPU.

4.1 Main Results

Table 1 compares baseline training with five clustering variants (DBSCAN, HDBSCAN, and K-means with $k \in \{1000, 2000, 5000\}$) across six metrics. We report Sentence-BLEU-1 for lexical overlap; COMET, which penalizes irrelevant or repetitive content and is therefore the metric most sensitive to noisy enrichment in our setting; BERTScore-F1 for contextual semantic similarity; and the clinically grounded CheXbert-F1 (disease-label agreement) and RadGraph-F1 (entity–relation-graph agreement). The pattern is consistent across all of them: HDBSCAN is the strongest clustering choice and has the best mean on every metric and K-means-5000 is a close second. Corpus-BLEU, METEOR, ChrF++, and ROUGE are deferred to Appendix A.12 and show the same qualitative trend; n-gram corpus-level scores favor longer outputs and, since our method adds sentences, should be read with caution, so we rely primarily on the semantic and clinical metrics here.

Although this work focuses on improving the training data rather than on producing the single best model, for completeness we situate our absolute numbers against open-source state-of-the-art medical vision–language models in Appendix A.6. These comparisons only locate where a report-only fine-tune lands relative to released checkpoints; the evidence for the enrichment itself comes from the controlled baseline-vs-enriched comparison (Table 1) and the size-matched random-expansion ablation (Section 4.2), not from the SOTA table.

Generalization to a second dataset: To test whether the enrichment transfers beyond ReXGradient-160K, we additionally run the full SemEnrich pipeline on MIMIC-CXR (Johnson et al., 2019). Despite MIMIC's single-institution origin and noisier reports, the pipeline transfers cleanly and HDBSCAN enrichment again improves a Mistral-7B backbone across lexical, semantic, and clinical metrics; the full setup and numbers are in Appendix A.7, and a comparison against medical VLMs on MIMIC is in Appendix A.8.

4.2 Ablation Study

To isolate the contribution of semantic clustering from that of simply adding more sentences, we compare three settings: the raw baseline, HDBSCAN-based enrichment (our method), and a *random-expansion* ablation in which we keep the same expansion size as our method but replace each added sentence with one drawn uniformly at random from a different cluster, ignoring connectivity and polarity.

As shown in Table 2, HDBSCAN enrichment improves over the baseline on every backbone and every metric, while random expansion is consistently *below* the baseline. The gain therefore stems from the semantic structure of the clusters, not from the additional sentence count alone.

4.3 GRPO Rewarding Results

We now evaluate the cluster-aware reward $\mathcal{R}_{cl}$ introduced in Section 3.4 against the standard exact-match reward $\mathcal{R}_{exact}$, in a GRPO stage that follows SFT. Recall that $\mathcal{R}_{cl}$ adds the F1 score between the ground-truth and generated cluster sets for the *findings* section on top of the exact-match reward for the *impression* section; we use HDBSCAN clustering throughout.

The results in Table 3 are computed on the impression section only. This is a deliberately conservative protocol: our reward augments the findings part, and we test whether the gain transfers to the impression – the final conclusion of

Table 1: Main results across five LLM backbones (mean±std over five seeds). Each block is one metric (in percent); the best method per backbone is bolded and the method with the best mean is shaded. Additional results are in Appendix A.12.

Metric	Method	DS-R1-Qwen-1.5B	Gemma3-4B	DS-R1-Llama-8B	Qwen3-8B	Mistral-7B	*Mean*
Sentence-BLEU-1	Baseline	27.58±0.87	21.63±0.70	27.45±0.86	28.11±1.17	29.87±1.67	26.93
	K-means-1000	26.40±0.66	22.80±0.74	26.85±1.00	26.88±0.71	23.12±0.52	25.21
	K-means-2000	28.74±0.97	22.25±0.70	29.09±0.77	29.36±0.94	29.58±0.69	27.80
	K-means-5000	28.16±0.62	23.32±0.32	**29.38±0.61**	29.40±0.79	32.53±1.15	28.56
	DBSCAN	27.03±1.31	21.36±0.83	28.44±0.69	28.02±0.56	28.89±1.00	26.75
	HDBSCAN	**29.51±0.81**	**23.83±0.91**	29.11±0.54	**30.35±1.35**	**32.64±1.00**	**29.09**
COMET	Baseline	63.68±0.49	62.05±0.53	63.95±0.48	63.80±0.80	63.59±0.77	63.41
	K-means-1000	65.76±0.45	63.65±0.39	66.66±0.37	66.32±0.40	65.42±0.26	65.56
	K-means-2000	67.14±0.43	63.51±0.32	67.37±0.37	67.40±0.49	68.10±0.40	66.70
	K-means-5000	66.38±0.49	63.97±0.22	67.02±0.36	67.16±0.52	67.33±0.51	66.37
	DBSCAN	63.28±0.65	61.93±0.35	64.24±0.59	64.09±0.63	63.65±0.60	63.44
	HDBSCAN	**67.34±0.41**	**64.59±0.65**	**67.44±0.31**	**67.71±0.73**	**69.16±0.57**	**67.25**
BERTScore-F1	Baseline	67.60±0.48	65.35±0.31	67.83±0.44	67.79±0.67	67.86±0.75	67.29
	K-means-1000	68.62±0.37	**66.93±0.41**	69.20±0.51	69.11±0.34	67.92±0.28	68.35
	K-means-2000	69.54±0.66	66.38±0.24	69.92±0.50	69.85±0.44	69.14±0.58	68.97
	K-means-5000	69.15±0.47	66.72±0.13	**69.97±0.39**	69.87±0.36	**71.40±0.59**	69.42
	DBSCAN	66.94±0.66	65.17±0.31	67.74±0.53	67.54±0.34	67.09±0.65	66.89
	HDBSCAN	**69.66±0.42**	66.80±0.60	69.80±0.34	**70.07±0.69**	71.35±0.56	**69.54**
CheXbert-14	Baseline	36.38±1.97	27.12±1.93	**41.34±2.37**	37.98±3.77	37.34±2.22	36.03
	K-means-1000	35.59±2.12	25.11±1.54	39.60±2.78	37.23±2.21	21.55±1.54	31.82
	K-means-2000	35.90±2.98	19.93±1.23	38.78±1.24	38.17±2.77	38.43±2.70	34.24
	K-means-5000	35.17±2.99	25.11±1.66	40.18±2.51	38.80±3.42	**39.30±3.31**	35.71
	DBSCAN	34.94±2.01	22.33±1.21	38.32±2.03	37.74±2.77	38.23±1.96	34.31
	HDBSCAN	**36.52±1.69**	**29.69±3.32**	39.72±2.38	**39.31±1.44**	39.25±2.71	**36.90**
RadGraph-F1	Baseline	26.51±1.00	18.49±1.07	26.08±1.18	27.32±1.69	30.21±1.72	25.72
	K-means-1000	24.93±1.03	20.62±0.82	26.07±1.40	25.97±1.35	23.69±1.02	24.26
	K-means-2000	**28.25±1.62**	18.90±0.93	**28.34±1.28**	28.69±1.47	31.45±1.47	27.13
	K-means-5000	26.04±1.00	19.83±0.97	27.92±0.99	28.32±1.22	31.65±1.72	26.75
	DBSCAN	24.03±1.60	16.23±1.22	25.16±1.04	24.78±1.14	28.73±1.61	23.79
	HDBSCAN	27.90±1.22	**20.71±1.55**	28.07±0.80	**29.28±2.00**	**32.23±1.60**	**27.64**

the report – which would indicate that the model has produced more informative findings as a by-product. Each run is averaged over 5 random seeds. Across both backbones and all three metrics, $\mathcal{R}_{cl}$ exceeds $\mathcal{R}_{exact}$ (+2.78% COMET, +3.14% BERTScore, +14.21% ROUGE-L on average). Additional GRPO details are in Appendix A.14.

Table 3: GRPO results comparing cluster-based reward $\mathcal{R}_{cl}$ and exact match reward $\mathcal{R}_{exact}$ on the impression section. Additional results are in Appendix A.14.

Metric	Model	$\mathcal{R}_{cl}$	$\mathcal{R}_{exact}$
COMET	Qwen2.5-VL-3B	**0.6179±0.003**	0.5973±0.002
	Qwen3-VL-4B	**0.6125±0.003**	0.5989±0.003
BERTScore-F1	Qwen2.5-VL-3B	**0.7211±0.002**	0.6952±0.002
	Qwen3-VL-4B	**0.7193±0.003**	0.7000±0.002
ROUGE-L	Qwen2.5-VL-3B	**0.3204±0.004**	0.2792±0.004
	Qwen3-VL-4B	**0.3203±0.005**	0.2818±0.005

5 Discussion

We presented SemEnrich, which turns the structure of a radiology corpus into a self-supervised training signal: clustering report sentences and reinserting plausibly omitted normal observations enriches the training data without any external annotation, and consistently improves report generation across five LLM backbones. The same clustering is also reusable as a reward signal for GRPO post-training, where a cluster-F1 reward complements the exact-match reward and yields further gains. Enrichment and the reward are orthogonal – one changes the training data by adding plausibly omitted normal sentences, while the other leaves the data untouched and instead reshapes the supervision signal at the cluster level.

Limitations. The sentences added through our enrichment method are not clinically verified by domain experts. Although the enrichment is constrained by connectivity and polarity conditions—ensuring that only *normal* findings from co-occurring clusters are appended—the added sentences are derived algorithmically from the training corpus rather than confirmed by a radiologist for

Table 2: Ablation: semantic (HDBSCAN) enrichment vs. random expansion (mean±std over five seeds, percent). The best method per backbone is bolded and the method with the best mean is shaded. Sentence-BLEU-1 and CheXbert-5 results are in Appendix A.12.

Metric	Method	DS-R1-Qwen-1.5B	Gemma3-4B	DS-R1-Llama-8B	Qwen3-8B	Mistral-7B	*Mean*
COMET	Baseline	63.68±0.49	62.05±0.53	63.95±0.48	63.80±0.80	63.59±0.77	63.41
	Random expansion	63.97±0.48	61.97±0.41	64.53±0.57	64.35±0.28	61.53±0.47	63.27
	HDBSCAN	**67.34±0.41**	**64.59±0.65**	**67.44±0.31**	**67.71±0.73**	**69.16±0.57**	**67.25**
BERTScore-F1	Baseline	67.60±0.48	65.35±0.31	67.83±0.44	67.79±0.67	67.86±0.75	67.29
	Random expansion	67.06±0.55	65.43±0.31	67.91±0.55	67.62±0.11	66.50±0.57	66.91
	HDBSCAN	**69.66±0.42**	**66.80±0.60**	**69.80±0.34**	**70.07±0.69**	**71.35±0.56**	**69.54**
RadGraph-F1	Baseline	26.51±1.00	18.49±1.07	26.08±1.18	27.32±1.69	30.21±1.72	25.72
	Random expansion	23.57±1.14	17.90±1.02	24.79±1.50	24.28±0.74	23.99±1.20	22.90
	HDBSCAN	**27.90±1.22**	**20.71±1.55**	**28.07±0.80**	**29.28±2.00**	**32.23±1.60**	**27.64**
CheXbert-14	Baseline	36.38±1.97	27.12±1.93	**41.34±2.37**	37.98±3.77	37.34±2.22	36.03
	Random expansion	30.04±1.32	23.12±2.55	28.85±2.38	33.63±3.77	25.15±1.38	28.16
	HDBSCAN	**36.52±1.69**	**29.69±3.32**	39.72±2.38	**39.31±1.44**	**39.25±2.71**	**36.90**

each individual patient. As a result, the enriched reports should not be treated as ground-truth clinical documents; they are intended exclusively as a training signal for machine learning models, and in a clinical deployment setting all generated reports would still require review and validation by qualified radiologists before any diagnostic use. Our approach also inherits the limitations of the underlying clustering and sign-assignment steps: the choice of sentence encoder, clustering algorithm, and number of clusters affects both the granularity of the enrichment and the noise in cluster membership, and our sign labels rely on an LLM judgment on a single representative sentence per cluster rather than a curated taxonomy.

References

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.

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

Coen Bron and Joep Kerbosch. 1973. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, and 1 others. 2024. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002*.

Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiaxiang Liu, and Zuozhu Liu. 2025. R-llava: Improving med-vqa understanding through visual region of interest. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

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

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, and 1 others. 2024. CheXagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M Sutter, Julia E Vogt, and 1 others. 2025. Radvlm: A multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and 1 others. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 2026. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *Advances in Neural Information Processing Systems*, 38.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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

Richard M Karp. 2009. Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*, pages 219–241. Springer.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. 2025. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, 35(7):4374–4386.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a.

Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

John W Moon and Leo Moser. 1965. On cliques in graphs. *Israel journal of Mathematics*, 3(1):23–28.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, and 1 others. 2025. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608.

Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. 2006. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical computer science*, 363(1):28–42.

Shuji Tsukiyama, Mikio Ide, Hiromu Ariyoshi, and Isao Shirakawa. 1977. A new algorithm for generating all the maximal independent sets. *SIAM Journal on Computing*, 6(3):505–517.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36:56186–56197.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiaoman Zhang, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

A Appendix

A.1 Cluster Examples

Additional semantic clusters, continuing from Cluster 1 in the main text.

Cluster 2: IVC Filter

"IVC filter noted within abdomen." / "IVC filter seen in the abdomen." / "IVC filter partially visualized in the abdomen." / ...

Cluster 3: Normal Size

"moderate size is normal." / "there is normal size." / "it is grossly normal in size." / "it is normal size and shape." / ...

Cluster 4: Imaging Artifacts

"this could be due to the artifacts." / "this may be an artifact." / "this could be an artifact." / "i suspect this is an artifact, but needs to be followed." / ...

Cluster 5: Pelvic Phleboliths

"bilateral pelvic phleboliths." / "pelvic calcifications consistent with phleboliths." / "multiple pelvic phleboliths." / "numerous pelvic phleboliths are noted." / ...

Cluster 6: Cerebral Infarcts

"focal encephalomalacia within the right occipital lobe...consistent with remote infarcts." / "there is a large left middle cerebral artery territory infarct..." / "there is an old left periventricular frontal lobe infarct." / ...

Cluster 7: Sickle Cell Disease Changes

"osseous changes of sickle cell disease are chronic and stable." / "bony changes of sickle cell disease are noted." / "biconcave vertebral body configuration is typical for sickle cell disease." / ...

A.2 Enrichment Examples

More examples of valid enrichments. We add a random sentence from an expansion cluster to the current sample during training. The first two boxes are alternative enrichments for the same initial findings used in the main-text example (Right hilar mass; heart normal; lungs clear; no pneumothorax or effusion; right-upper-lobe airspace disease), illustrating that one report can admit several distinct valid enrichments.

Alternative Enrichment A (Same Initial Findings)

Expansion Cluster 1: *"no bony abnormality is seen", "no acute bony abnormalities", "no acute bony abnormality is seen", ...*
Expansion Cluster 2: *"normal lung volumes", "lung volumes are normal", "the lung volumes are normal", ...*

Alternative Enrichment B (Same Initial Findings)

Expansion Cluster 1: *"trachea is midline", "the trachea is midline", "midline trachea", ...*
Expansion Cluster 2: *"visualized bony thorax intact", "the visualized bony structures of the thorax are intact", ...*

Enrichment Example 1

Initial Findings:

- The heart size and mediastinal contours are within normal limits.
- Both lungs are clear.
- The visualized skeletal structures are unremarkable.
- Bilateral nipple rings.

Expansion Cluster: *"there is no pleural effusion or edema identified", "no pleural effusions", "no pleural effusion", ...*

Enrichment Example 2

Initial Findings:

- Heart: Normal. Lungs: The lungs are hyperinflated.
- There is mild scattered chronic lung changes throughout both lung fields, most pronounced in the right apex.
- There is a 8 mm calcified granuloma in the left lung base. Otherwise, the lung fields are relatively clear.
- Mediastinum: Normal. Pleural effusions: None.
- There is suggestion of diffuse osteopenia. Correlate for osteoporosis.
- There is moderate-to-marked chronic anterior wedge compression deformity of the approximately T9 vertebral body.

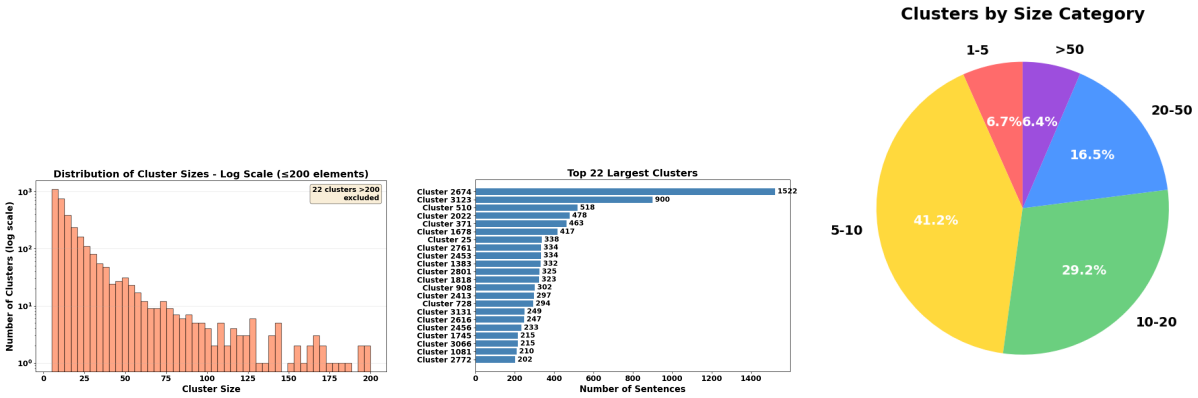
Expansion Cluster: *"no pneumothorax is seen", "there is no pneumothorax", "no definite pneumothorax is seen", ...*

A.3 HDBSCAN Clustering Statistics

Although we have 184,535 sentences in the training dataset, as HDBSCAN allows outliers and not to assign them to a cluster, eventuall we end up with 66,477 sentences in the training dataset that are assigned to a cluster. This is one of the main difference between density based clustering and K-means.

Statistic	Value
Total Sentences	66,477
Total Clusters	3,216
Average Cluster Size	20.67
Median Cluster Size	11.0
Min Cluster Size	5
Max Cluster Size	1,522

Table 4: HDBSCAN clustering statistics for finding sentences.



(a) (b) (c)

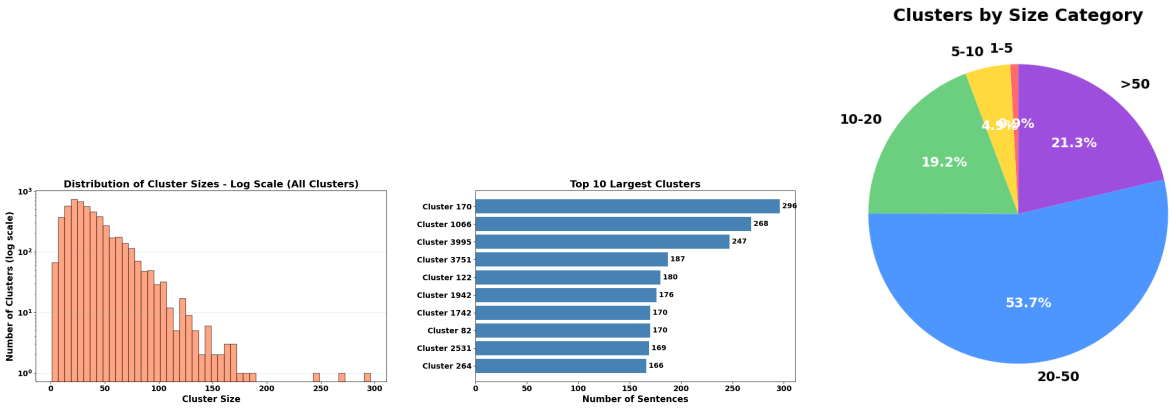
Figure 6: HDBSCAN clustering statistics: (a) Log histogram of cluster sizes, (b) Top clusters by size, (c) Percentage distribution of clusters.

A.4 K-means Clustering Statistics with 5000 Clusters

We provide the details of the K-means clustering with 5000 clusters as an example to illustrate the clustering difference between density based clustering methods and K-means.

Statistic	Value
Total Sentences	184,535
Total Clusters	5,000
Average Cluster Size	36.91
Median Cluster Size	31.0
Min Cluster Size	1
Max Cluster Size	296

Table 5: K-means-5000 clustering statistics for finding sentences.



(a) (b) (c)

Figure 7: K-means-5000 clustering statistics: (a) Log histogram of cluster sizes, (b) Top clusters by size, (c) Percentage distribution of clusters.

A.5 Largest Valid Enrichment Problem

The problem of finding all maximal expansions in our framework is closely related to the classical *maximal clique enumeration* problem in graph theory. Given an undirected graph $G = (V, E)$, a *clique* is a subset of vertices in which every pair is connected by an edge. A clique is *maximal* if no additional vertex can be included without breaking the clique property, and it is *maximum* if it has the largest cardinality among all cliques in the graph. The decision version of the maximum clique problem—determining whether a clique of size at least k exists—is one of Karp's original 21 NP-complete problems (Karp, 2009). Moon and Moser (Moon and Moser, 1965) established that a graph on n vertices can contain up to $3^{n/3}$ maximal cliques, and this bound is tight. Consequently, any algorithm that enumerates all maximal cliques requires exponential time in the worst case simply because the output itself can be exponentially large.

The most widely used algorithm for maximal clique enumeration is the Bron–Kerbosch algorithm (Bron and Kerbosch, 1973), a backtracking procedure that maintains three sets: R (the current clique), P (candidate vertices that can extend R), and X (vertices already processed that ensure maximality). With the *pivoting* optimization (Tomita et al., 2006), the algorithm achieves $O(3^{n/3})$ worst-case time, matching the Moon–Moser bound and making it output-optimal. Furthermore, Tsukiyama et al. (Tsukiyama et al., 1977) showed that maximal cliques can be enumerated with *polynomial delay*—the time between outputting consecutive cliques is bounded by a polynomial in n.

In our setting, finding all largest valid enrichments for a patient with findings F_j corresponds to enumerating all maximal cliques in the subgraph induced by $\mathcal{N}^+(F_j)$: the set of normal clusters that are addable to every cluster in F_j. Since $|\mathcal{N}^+(F_j)|$ is typically much smaller than the total number of clusters $|C|$—due to the intersection of neighborhoods—the enumeration remains tractable in practice. We precompute all maximal expansions for every unique finding graph before training, and during training we uniformly sample one expansion per instance in each batch.

The addability matrix B in Algorithm 1 depends on two thresholds, which we fix across all configurations: $\tau_{\text{norm}} = 0.001$, a floor on the row-normalized co-occurrence $\tilde{A}[i, j]$ between two clusters, and $\tau_{\text{count}} = 3$, a floor on the raw number of reports $A[i, j]$ in which they co-occur. A candidate normal cluster is addable to a report only if it clears both thresholds with respect to every cluster already present, in addition to passing the polarity gate ($h_S(j) = +1$).

Algorithm 1 Precomputing All Maximal Valid Enrichments

Require: Training data $\mathcal{D} = \{(v_i, F_i)\}_{i=1}^{N}$, cluster mapping f_C, cluster signs h_S, thresholds $\tau_{\text{norm}}, \tau_{\text{count}}$
Ensure: Expansion dictionary $\mathcal{E}$

// Stage 1: Build Co-occurrence and Addability
Initialize $A \in \mathbb{R}^{K \times K} \leftarrow \mathbf{0}$ ▷ K = number of clusters
for each $(v_i, F_i) \in \mathcal{D}$ **do**
  $G_i \leftarrow \{f_C(s) : s \in F_i\}$
  for each pair $(c_a, c_b) \in G_i \times G_i$, $c_a \neq c_b$ **do**
    $A[c_a, c_b] \leftarrow A[c_a, c_b] + 1$
  end for
end for
$\tilde{A} \leftarrow \text{RowNormalize}(A)$
$B[i, j] \leftarrow \mathbf{1}\Big[\tilde{A}[i, j] > \tau_{\text{norm}} \ \wedge \ A[i, j] > \tau_{\text{count}} \ \wedge \ h_S(j) = +1\Big]$

// Stage 2: Enumerate Maximal Expansions
for each unique finding graph G **do**
  $\mathcal{N}^+(G) \leftarrow \bigcap_{c \in G} \{j : B[c, j] = 1\} \setminus G$ ▷ Compatible normal clusters
  $\mathcal{E}[G] \leftarrow \textsc{Expand}(G, \ \text{sorted}(\mathcal{N}^+(G)), \ 0)$
  Remove any $R_i \in \mathcal{E}[G]$ s.t. $\exists\, R_j \in \mathcal{E}[G]$: $R_i \subset R_j$ ▷ Keep only maximal
end for
return $\mathcal{E}$

Function $\textsc{Expand}(R, \ P, \ s)$
expanded $\leftarrow$ false
for $i = s$ **to** $|P|$ **do**
  if $B[c, \ p_i] = 1$ for all $c \in R$ **then**
    $\textsc{Expand}(R \cup \{p_i\}, \ P, \ i + 1)$
    expanded $\leftarrow$ true
  end if
end for
if not expanded **then**
  emit R ▷ No candidate can extend R
end if
end function

Table 6: Comparison with open-source state-of-the-art medical VLMs on the ReXGradient-160K validation split (2,500 reports). SOTA rows are single inference runs produced by each model's released checkpoint; Mistral-7B rows are mean±std over five training seeds (taken from Table 1). RadGraph-F1 is the *complete*-match variant for all rows. Best value per metric is bolded. All scores are in percent.

Model	COMET	BERTScore	S-BLEU-1	RadGraph
Med-Flamingo	51.18	62.63	6.79	10.47
LLaVA-Med	61.31	58.17	8.80	4.44
CXR-LLaVA	67.40	63.56	17.60	13.63
CheXagent	60.81	63.46	17.66	14.83
LLaVA-Rad	65.75	65.72	18.57	15.75
Mistral-7B (Baseline)	63.59±0.77	67.86±0.75	29.87±1.67	30.21±1.72
Mistral-7B (HDBSCAN, ours)	**69.16±0.57**	**71.35±0.56**	**32.64±1.00**	**32.23±1.60**

A.6 Comparison with State-of-the-Art Medical VLMs

We benchmark our enriched Mistral-7B backbone against five open-source medical vision–language models on the ReXGradient-160K validation split (2,500 reports): Med-Flamingo (Moor et al., 2023), LLaVA-Med (Li et al., 2023a), CXR-LLaVA (Lee et al., 2025), CheXagent (Chen et al., 2024), and LLaVA-Rad (Chaves et al., 2024). SOTA outputs are produced by each model's released checkpoint and evaluated with the same metric pipeline used for our backbones (in particular, RadGraph-F1 uses the *complete*-match variant for both our methods and SOTA, matching the value reported for Mistral-7B in Table 1). SOTA numbers are therefore single inference runs (no seed variance), while our Mistral-7B rows are mean±std over five training seeds and are reproduced from Table 1.

We stress that this is *not* a matched comparison and is included only to situate our absolute numbers: the SOTA models are evaluated zero-shot from released checkpoints and were never trained on ReXGradient-160K, whereas our backbone is fine-tuned in-domain. The table therefore mostly reflects the benefit of in-domain report-only fine-tuning, not the enrichment; the controlled evidence for enrichment is the baseline-vs-enriched comparison (Table 1) and the random-expansion ablation (Table 2).

Table 6 reports the four shared metrics. Our HDBSCAN-enriched Mistral-7B is the top model on all four: COMET (+1.8 over the strongest SOTA), BERTScore-F1 (+5.6), Sentence-BLEU-1 (+14.1), and RadGraph-F1 (+16.5). Indeed the unenriched Mistral-7B baseline already matches or exceeds every SOTA model on each metric, which confirms that the gap chiefly reflects in-domain fine-tuning; semantic enrichment then improves further on top of that already-strong baseline.

A.7 Generalization to MIMIC-CXR

To assess whether SemEnrich transfers beyond ReXGradient-160K, we run the complete pipeline—sentence embedding, clustering, polarity labelling, connectivity, and enrichment—on the MIMIC-CXR dataset (Johnson et al., 2019), using a 20K-report training subset. The pipeline transfers without modification: HDBSCAN produces 836 clusters, and enrichment touches 58.8% of reports, adding on average 1.2 sentences per enriched report. For polarity labelling on MIMIC we use a 3-sample majority vote per cluster to reduce the single-sentence labelling noise discussed in the main text. We then fine-tune a Mistral-7B backbone on the raw and enriched data under the same recipe as our main experiments and report the mean±std over four seeds.

As shown in Table 7, HDBSCAN enrichment improves the backbone on every metric, including the clinically grounded CheXbert-14 (+3.10) and RadGraph-F1 (+2.59). MIMIC is a single-institution corpus with noisier reports than ReXGradient-160K (Appendix A.15), so the fact that the enrichment still helps is evidence that it is not specific to one dataset or reporting style.

Table 7: Generalization of SemEnrich to MIMIC-CXR (20K training reports). Mistral-7B backbone, mean±std over four seeds; all scores in percent. Δ is HDBSCAN minus Baseline.

Metric	Baseline	HDBSCAN (ours)	Δ
Sentence-BLEU-1	23.88±0.91	**30.35±0.18**	+6.47
METEOR	27.04±0.99	**30.17±0.32**	+3.13
CheXbert-14	52.70±0.44	**55.80±2.24**	+3.10
RadGraph-F1	22.69±1.17	**25.28±0.72**	+2.59

A.8 Comparison with Medical VLMs on MIMIC-CXR

For completeness we also situate the MIMIC numbers against the same open-source medical vision–language models used in Appendix A.6. As before, this is *not* a matched comparison: the SOTA models are single zero-shot inference runs from released checkpoints, and it is included only to locate our absolute performance. Note that two of these models, CheXagent (Chen et al., 2024) and LLaVA-Rad (Chaves et al., 2024), were themselves trained on MIMIC-CXR, so the setting is if anything favourable to them.

Even so (Table 8), our HDBSCAN-enriched Mistral-7B attains the best RadGraph-F1 and the second-best CheXbert-14, and improves over its own baseline on all four metrics. The lexical metrics (Sentence-BLEU-1, METEOR) still favour the MIMIC-trained models, consistent with the domain match.

Table 8: Comparison with open-source medical VLMs on MIMIC-CXR. SOTA rows are single inference runs from released checkpoints; our rows are the Mistral-7B means from Table 7. CheXagent and LLaVA-Rad were trained on MIMIC-CXR. Best value per metric is bolded; all scores in percent.

Model	S-BLEU-1	METEOR	CheXbert-14	RadGraph
Med-Flamingo (Moor et al., 2023)	13.05	13.62	54.00	18.59
LLaVA-Med (Li et al., 2023a)	17.99	14.54	54.00	7.68
CXR-LLaVA (Lee et al., 2025)	30.30	28.10	32.75	17.50
CheXagent (Chen et al., 2024)	35.69	32.31	**58.00**	22.35
LLaVA-Rad (Chaves et al., 2024)	**37.84**	**33.64**	55.25	19.35
Ours (Baseline)	23.88	27.04	52.70	22.69
Ours (HDBSCAN)	30.35	30.17	55.80	**25.28**

A.9 Sentence-Encoder Ablation

Since embedding quality determines cluster quality, which in turn determines which sentences are eligible for enrichment, we ablate the sentence encoder with the rest of the pipeline held fixed. We compare our default `all-MiniLM-L6-v2` against two larger general-purpose encoders (Stella, GTE-Qwen2) and two biomedical encoders (PubMedBERT, BioBERT). All runs use a Mistral-7B backbone on MIMIC-CXR.

Table 9 shows that *every* encoder improves over the baseline on all four metrics. The larger and biomedical encoders raise the lexical scores (Sentence-BLEU-1, METEOR), but the general-purpose MiniLM remains the strongest on both clinical metrics (CheXbert-14 and RadGraph-F1), and the biomedical encoders do not beat it. Finally, the difference in cluster coverage is not an embedding-quality artifact: the larger and biomedical encoders assign *fewer* sentences to clusters (e.g., MiniLM 16,961 vs. PubMedBERT 6,855 and BioBERT 8,179), and sentences left unclustered still remain in their reports—they are only excluded from the candidate enrichment pool, not removed from training. We therefore keep MiniLM, also the top-listed general-purpose model in the Sentence-Transformers library, as the default.

Table 9: Sentence-encoder ablation with the pipeline held fixed (Mistral-7B, MIMIC-CXR). "Clust./Assig." is the number of clusters and the number of sentences assigned to a cluster. Best value per metric is bolded; scores in percent.

Encoder (dim)	S-BLEU-1	METEOR	CheXbert-14	RadGraph	Clust./Assig.
Baseline	23.88	27.04	52.70	22.69	—
MiniLM (384)	30.35	30.17	**55.80**	**25.28**	836 / 16,961
Stella (1024)	35.70	33.41	54.40	23.28	534 / 12,772
GTE-Qwen2 (1536)	33.76	31.97	52.90	22.05	576 / 11,399
PubMedBERT (768)	**36.59**	**35.88**	52.80	24.44	536 / 6,855
BioBERT (768)	24.75	27.70	54.10	23.13	628 / 8,179

A.10 Which Normal Findings Are Recovered

To show that the enrichment repairs a systematic reporting bias rather than merely moving a metric, we inspect what the procedure actually inserts. On MIMIC-CXR, of the 836 clusters, 133 are labelled normal and 703 abnormal; enrichment touches 58.8% of reports ($+1.2$ sentences each), and by construction every inserted cluster is polarity-normal. The normal findings that are most frequently recovered (i.e. most often omitted from the original report) are the canonical ones, as shown in Table 10.

Table 10: Most frequently recovered normal findings on MIMIC-CXR, as a percentage of training reports into which they are inserted.

Recovered normal finding	% of reports
normal heart size	22.0
no osseous abnormalities	21.8
no pleural effusion or pneumothorax	19.6
no pneumothorax	11.0
normal pulmonary vasculature	8.2

Aligning baseline and enriched generations exposes the mechanism directly: the enriched model adds omitted normals (e.g., "hilar contours unremarkable"), which helps when the reference report states them and lowers overlap only when the reference is terse. In the latter case the drop reflects the metric penalising completeness rather than a genuine error, consistent with the enrichment making implicit normal findings explicit.

A.11 Within-Cluster Polarity Purity

Because only clusters labelled *normal* are ever inserted, what matters for label noise is the polarity purity of those clusters. We therefore LLM-labelled $\sim 1{,}000$ member sentences across all 133 inserted normal clusters on MIMIC-CXR (for large clusters we sample up to 10 members) and measured how consistently each cluster's members share its assigned polarity. Table 11 shows the clusters are highly pure.

Table 11: Polarity purity of the 133 inserted normal clusters (MIMIC-CXR). Insertion-weighted purity weights each cluster by how often it is inserted during training; the abnormal-injection rate is its complement.

Statistic	%
Mean purity	97.9
Median purity	100.0
Clusters 100% pure	92.5
Clusters $\geq 80\%$ pure	97.0
Insertion-weighted purity	99.78
Abnormal-injection rate	0.22

Weighting each cluster by how often it is actually inserted during training, the clusters we add are $\sim 99.8\%$ pure, corresponding to an effective abnormal-injection rate of only $\sim 0.22\%$; in other words, the procedure essentially never introduces an abnormal finding. The few genuinely mixed clusters are rare and low-frequency, so a conservative safeguard is nearly free: dropping every cluster below 80% purity would remove just 0.27% of the total insertion mass. As noted above, our MIMIC runs additionally use a 3-sample majority vote for the cluster sign to further reduce labelling noise.

A.12 Semantic Enrichment Additional Results

We first report the CheXbert-5 F1 block deferred from the main results (Table 12); the corresponding CheXbert-5 ablation block is reported in the Ablation Study Results section below.

Table 12: CheXbert-5 F1 scores comparing baseline with clustering methods (deferred from the main results).

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	17.26±4.85	18.37±2.99	21.19±5.09	**23.54±4.17**	23.23±4.84	21.65±4.93
Gemma3-4B	17.73±1.85	9.34±2.77	**20.06±2.05**	17.17±4.32	11.07±2.18	12.88±1.16
DS-R1-Llama-8B	**31.90±3.44**	23.15±2.27	24.93±4.47	29.70±2.37	26.30±3.89	27.51±5.28
Qwen3-8B	21.97±5.39	22.48±4.14	25.02±3.51	**26.22±3.12**	21.78±4.80	25.77±4.69
Mistral-7B	15.89±4.30	17.01±4.22	19.13±5.79	**24.82±3.76**	16.03±3.01	21.67±2.87
Mean	20.95	18.07	22.06	**24.29**	19.68	21.90

We additionally provide enrichment results for METEOR, ChrF++, ROUGE, and BERTScore (using RoBERTa-large as the underlying encoder).

Table 13: METEOR scores comparing baseline with clustering methods

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	26.36±0.98	26.97±1.70	30.13±1.18	29.90±1.27	**30.94±1.53**	28.63±1.20
Gemma3-4B	20.48±0.64	20.70±0.98	23.52±1.16	**25.68±0.95**	23.18±0.64	23.04±0.08
DS-R1-Llama-8B	26.26±1.12	28.10±1.21	29.72±0.79	**30.87±1.43**	30.51±1.05	30.46±1.27
Qwen3-8B	27.21±1.43	27.57±0.89	31.21±1.97	29.92±1.37	**31.57±1.36**	30.38±0.86
Mistral-7B	29.42±1.83	27.87±1.52	33.51±1.66	30.67±1.09	33.13±1.41	**33.78±1.63**
Mean	25.95	26.24	29.62	29.41	**29.87**	29.26

Table 14: ChrF++ scores comparing baseline with clustering methods

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	35.39±0.81	36.90±1.48	39.40±1.04	40.43±0.94	**40.88±1.41**	38.72±1.03
Gemma3-4B	31.61±0.66	32.40±0.71	34.50±1.15	**37.03±0.75**	34.91±0.57	34.34±0.09
DS-R1-Llama-8B	35.33±0.88	38.13±1.07	38.99±0.75	**41.42±1.22**	40.83±1.01	40.03±1.18
Qwen3-8B	35.97±1.43	37.66±0.67	39.99±1.60	40.69±0.91	**41.26±1.15**	39.85±0.81
Mistral-7B	37.48±1.65	38.85±1.41	42.67±1.42	41.83±0.77	43.23±1.32	**43.25±1.50**
Mean	35.16	36.79	39.11	40.28	**40.22**	39.24

Table 15: Corpus BLEU-1 scores comparing baseline with clustering methods (Pretrain).

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	25.59±0.95	29.53±1.42	**32.44±0.96**	27.42±0.67	31.05±0.65	31.98±0.86
Gemma3-4B	23.67±0.69	24.77±1.05	**27.64±0.95**	23.19±0.71	23.72±0.74	26.93±0.49
DS-R1-Llama-8B	26.73±1.17	31.19±1.30	32.02±0.88	28.06±1.19	31.86±0.89	**33.17±0.91**
Qwen3-8B	26.14±1.21	29.93±0.87	33.06±1.37	28.45±0.69	32.09±0.90	**33.19±0.60**
Mistral-7B	25.44±1.51	29.89±1.02	**34.72±1.29**	23.76±0.57	32.30±0.74	34.71±1.57
Mean	25.51	29.06	**31.97**	26.18	30.21	31.99

Table 16: ROUGE-1-F scores comparing baseline with clustering methods (Pretrain).

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	37.79±0.99	36.13±1.34	**40.54±0.99**	37.53±0.75	39.83±1.20	38.96±0.86
Gemma3-4B	31.59±0.86	30.52±0.96	**34.09±1.21**	33.51±0.71	32.43±0.44	33.32±0.17
DS-R1-Llama-8B	37.80±0.81	37.80±0.83	**40.47±0.79**	38.27±1.04	40.24±0.93	40.39±0.85
Qwen3-8B	38.35±1.30	37.09±0.59	**41.40±1.52**	38.10±0.69	40.36±0.99	40.42±0.93
Mistral-7B	39.44±1.68	36.86±1.19	**44.12±1.15**	34.75±0.62	40.53±0.96	43.86±1.37
Mean	37.00	35.68	**40.12**	36.43	38.68	39.39

Table 17: ROUGE-2-F scores comparing baseline with clustering methods (Pretrain).

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	**19.04±0.86**	17.45±1.57	19.04±1.09	15.83±1.00	18.48±1.44	17.21±0.96
Gemma3-4B	12.06±0.51	10.05±0.75	**12.82±0.87**	11.55±0.68	10.55±0.68	11.87±0.41
DS-R1-Llama-8B	18.76±0.84	18.39±1.06	18.58±0.96	16.63±1.33	18.24±1.02	**18.96±1.04**
Qwen3-8B	**20.27±1.74**	18.22±0.65	20.10±1.75	16.29±1.03	18.99±1.38	18.74±0.99
Mistral-7B	**23.53±1.87**	20.96±1.57	22.56±1.53	15.82±0.82	21.12±1.26	23.34±1.43
Mean	**18.73**	17.01	18.62	15.22	17.48	18.02

Table 18: ROUGE-L-F scores comparing baseline with clustering methods (Pretrain).

Model	Baseline	DBSCAN	HDBSCAN	K-means-1000	K-means-2000	K-means-5000
DS-R1-Qwen-1.5B	30.32±0.89	28.32±1.22	**30.49±1.08**	26.60±0.72	29.39±1.07	28.79±0.57
Gemma3-4B	23.64±0.52	21.89±0.66	**24.67±0.88**	22.54±0.69	22.27±0.55	23.85±0.42
DS-R1-Llama-8B	29.96±0.76	29.40±0.98	30.32±0.84	27.17±0.97	29.48±0.76	**30.45±0.82**
Qwen3-8B	31.06±1.22	29.22±0.64	**31.27±1.33**	26.99±0.84	29.89±1.05	30.14±0.65
Mistral-7B	33.43±1.51	30.70±1.21	33.65±1.23	24.56±0.56	31.30±1.01	**34.26±1.15**
Mean	29.68	27.91	**30.08**	25.57	28.47	29.50

A.13 Ablation Study Results

We provided additional results for the ablation study below. For some metrics, although we observe an improvement from the random expansion due to metric hacking, we also see that semantically guided expansion yields additional improvements over the random expansion.

We first report the CheXbert-5 F1 ablation block deferred from the main paper.

Table 19: Ablation study: CheXbert-5 F1 comparing Baseline, HDBSCAN, and Random expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	17.26±4.85	**21.19±5.09**	20.86±5.00
Gemma3-4B	17.73±1.85	**20.06±2.05**	18.98±5.45
DS-R1-Llama-8B	**31.90±3.44**	24.93±4.47	21.68±4.60
Qwen3-8B	21.97±5.39	25.02±3.51	**26.82±5.30**
Mistral-7B	15.89±4.30	**19.13±5.79**	14.84±2.02
Mean	20.95	**22.06**	20.63

Table 20: Ablation study: COMET comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	63.68±0.49	**67.34±0.41**	63.97±0.48
Gemma3-4B	62.05±0.53	**64.59±0.65**	61.97±0.41
DS-R1-Llama-8B	63.95±0.48	**67.44±0.31**	64.53±0.57
Qwen3-8B	63.80±0.80	**67.71±0.73**	64.35±0.28
Mistral-7B	63.59±0.77	**69.16±0.57**	61.53±0.47
Mean	63.41	**67.25**	63.27

Table 21: Ablation study: Sentence BLEU-1 comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	27.58±0.87	**29.51±0.81**	26.88±1.01
Gemma3-4B	21.63±0.70	**23.83±0.91**	22.18±0.49
DS-R1-Llama-8B	27.45±0.86	**29.11±0.54**	27.59±0.71
Qwen3-8B	28.11±1.17	**30.35±1.35**	27.17±0.44
Mistral-7B	29.87±1.67	**32.64±1.00**	28.85±1.02
Mean	26.93	**29.09**	26.53

Table 22: Ablation study: METEOR comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	26.36±0.98	**30.13±1.18**	28.33±1.41
Gemma3-4B	20.48±0.64	**23.52±1.16**	22.24±0.61
DS-R1-Llama-8B	26.26±1.12	**29.72±0.79**	28.74±1.24
Qwen3-8B	27.21±1.43	**31.21±1.97**	28.34±0.60
Mistral-7B	29.42±1.83	**33.51±1.66**	32.33±1.72
Mean	25.95	**29.62**	28.00

Table 23: Ablation study: ROUGE-1-F comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	37.79±0.99	**40.54±0.99**	36.73±1.11
Gemma3-4B	31.59±0.86	**34.09±1.21**	31.77±0.59
DS-R1-Llama-8B	37.80±0.81	**40.47±0.79**	37.79±0.86
Qwen3-8B	38.35±1.30	**41.40±1.52**	37.28±0.44
Mistral-7B	39.44±1.68	**44.12±1.15**	38.24±1.12
Mean	37.00	**40.12**	36.36

Table 24: Ablation study: ROUGE-2-F comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	**19.04±0.86**	19.04±1.09	17.42±1.20
Gemma3-4B	12.06±0.51	**12.82±0.87**	11.95±0.81
DS-R1-Llama-8B	**18.76±0.84**	18.58±0.96	18.21±1.32
Qwen3-8B	**20.27±1.74**	20.10±1.75	17.82±0.45
Mistral-7B	**23.53±1.87**	22.56±1.53	20.36±1.32
Mean	**18.73**	18.62	17.15

Table 25: Ablation study: ROUGE-L-F comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	30.32±0.89	**30.49±1.08**	28.04±0.89
Gemma3-4B	23.64±0.52	**24.67±0.88**	23.28±0.70
DS-R1-Llama-8B	29.96±0.76	**30.32±0.84**	29.19±0.95
Qwen3-8B	31.06±1.22	**31.27±1.33**	28.55±0.42
Mistral-7B	33.43±1.51	**33.65±1.23**	29.57±1.08
Mean	29.68	**30.08**	27.72

Table 26: Ablation study: ChrF++ comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	35.39±0.81	**39.40±1.04**	37.56±1.25
Gemma3-4B	31.61±0.66	**34.50±1.15**	33.39±0.73
DS-R1-Llama-8B	35.33±0.88	**38.99±0.75**	37.80±1.38
Qwen3-8B	35.97±1.43	**39.99±1.60**	37.63±0.49
Mistral-7B	37.48±1.65	**42.67±1.42**	40.02±1.46
Mean	35.16	**39.11**	37.28

Table 27: Ablation study: BERTScore-F1 comparing Baseline, HDBSCAN, and Random Expansion.

Model	Baseline	HDBSCAN	Random expansion
DS-R1-Qwen-1.5B	67.60±0.48	**69.66±0.42**	67.06±0.55
Gemma3-4B	65.35±0.31	**66.80±0.60**	65.43±0.31
DS-R1-Llama-8B	67.83±0.44	**69.80±0.34**	67.91±0.55
Qwen3-8B	67.79±0.67	**70.07±0.69**	67.62±0.11
Mistral-7B	67.86±0.75	**71.35±0.56**	66.50±0.57
Mean	67.29	**69.54**	66.91

A.14 GRPO Training Additional Results

As we pointed out in the main paper, we observe significant performance improvements by using $\mathcal{R}_{cl}$ in the reward calculation of GRPO training. We provide results for different metrics below.

Table 28: Sentence-BLEU-1 scores, comparing $\mathcal{R}_{cl}$ and $\mathcal{R}_{exact}$ rewards (GRPO) on the impression section.

Model	$\mathcal{R}_{cl}$	$\mathcal{R}_{exact}$
Qwen2.5-VL-3B	**0.2704±0.0035**	0.2361±0.0042
Qwen3-VL-4B	**0.2706±0.0044**	0.2354±0.0034

Table 29: METEOR scores, comparing $\mathcal{R}_{cl}$ and $\mathcal{R}_{exact}$ rewards (GRPO).

Model	$\mathcal{R}_{cl}$	$\mathcal{R}_{exact}$
Qwen2.5-VL-3B	**0.2621 ± 0.0034**	0.2335 ± 0.0054
Qwen3-VL-4B	**0.2620 ± 0.0048**	0.2344 ± 0.0032

Table 30: Sentence-BLEU-4 scores, comparing $\mathcal{R}_{cl}$ and $\mathcal{R}_{exact}$ rewards (GRPO).

Model	$\mathcal{R}_{cl}$	$\mathcal{R}_{exact}$
Qwen2.5-VL-3B	**0.1825 ± 0.0029**	0.1599 ± 0.0027
Qwen3-VL-4B	**0.1847 ± 0.0041**	0.1578 ± 0.0023

A.15 On the Dataset

As noted in the main text, we use the ReXGradient-160K dataset for our experiments (Zhang et al., 2025). This dataset comprises 160K chest X-ray studies covering 109,487 unique patients from 79 medical sites. Since the data originates from 79 distinct sites, it exhibits considerable diversity in imaging conditions, acquisition protocols, and equipment. Moreover, the reports are authored by different radiologists, introducing substantial variability in writing style and clinical language.

We did not use the CheXpert dataset (Irvin et al., 2019) for our experiments for two reasons. First, it is sourced solely from Stanford Hospital and therefore may lack the multi-site diversity present in ReXGradient-160K. Second, it is structured more as a classification benchmark than a free-text report generation dataset, which may not be representative of real-world clinical reporting scenarios.

We use the MIMIC-CXR dataset (Johnson et al., 2019) not as our primary benchmark but as an out-of-distribution stress test of the enrichment (Appendix A.7). MIMIC is collected from a single medical center, which limits its institutional diversity relative to ReXGradient-160K, and its reports are noisier in the sense that findings or impressions may be missing or poorly structured—precisely the conditions under which our enrichment might be expected to struggle. That the full pipeline nonetheless transfers and improves a backbone on MIMIC is therefore stronger evidence for its generality than a second clean corpus would be.

More broadly, single-institution datasets limit the evaluation of a model's generalization capability across diverse healthcare settings. For these reasons, we chose the ReXGradient-160K dataset, which not only spans 79 medical centers but also ensures that every sample contains both findings and impressions, making it well suited for evaluating our method.

A.16 On the Enriched Dataset

It is important to note that the sentences added through our enrichment method are not clinically verified by domain experts. Although the enrichment is constrained by connectivity and polarity conditions—ensuring that only *normal* findings from co-occurring clusters are appended—the added sentences are derived algorithmically from the training corpus rather than confirmed by a radiologist for each individual patient.

As a result, the enriched reports should not be treated as ground-truth clinical documents. Instead, the enriched dataset is intended exclusively as a training signal for machine learning models. The goal is to provide the model with richer supervisory information during training by making implicit normal findings explicit, thereby encouraging the model to learn a more complete understanding of chest X-ray images.

We emphasize that this enrichment strategy is a data augmentation technique for model training and is not a substitute for expert annotation. In a clinical deployment setting, all generated reports would still require review and validation by qualified radiologists before any diagnostic use.