Summary

The more result is organized as follows:

- Section A.1 describes the prototype formulation procedure for our Pathology-Aligned Prototype Attention (PAPA) mechanism.
- Section A.2 provides complete architectural specifications for all model components.
- Section A.3 details training hyperparameters, optimization schedules, and computational setup.
- Section A.4 presents the Sinkhorn-Knopp algorithm for optimal transport-based prototype assignment.
- Section A.5 describes the patch-based masking methodology for visual grounding evaluation.
- Section B.1 describes the metrics for report generation.
- Section B.2 extends the visual masking experiments with comprehensive quantitative results across all metrics.
- Section B.3 provides detailed per-pathology clinical efficacy results and precision-recall trade-off analysis.

A. Implementation Details

This section provides additional implementation details that complement Section 4.2 (Experiment Settings) of the main paper.

A.1. Prototype Formulation

This subsection details the prototype formulation procedure mentioned in Section 3.3 (PAPA) of the main paper. The complete pipeline is illustrated in Figure S-I.

As described in Section 3.3, we define a set of K pathology prototypes $P = \{p_1, p_2, \ldots, p_K\} \in \mathbb{R}^{K \times D}$. To construct these prototypes with semantically meaningful representations, we follow a systematic procedure based on clinical pathology labels from the CheXpert dataset.

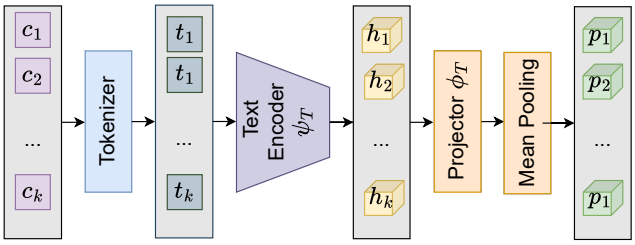

Figure S-I. **Pathology prototype formulation pipeline.** Given K pathology concept names $\{c_1, c_2, \ldots, c_K\}$ from CheXpert labels, each concept is tokenized into sequences $\{t_1, \ldots, t_k\}$, encoded by the text encoder ψ_T to obtain contextualized embeddings $\{h_1, \ldots, h_k\}$, projected into the shared latent space via ϕ_T, and finally mean-pooled across the sequence dimension to produce fixed prototype vectors $\{p_1, p_2, \ldots, p_K\}$ that serve as semantic anchors for pathology-aware alignment.

Pathology Concepts. Following CheXpert, which defines a standard set of 14 clinical findings for chest X-ray analysis, we use $K = 14$ pathology concepts for prototype construction: *"enlarged cardiomediastinum"*, *"cardiomegaly"*, *"lung opacity"*, *"lung lesion"*, *"edema"*, *"consolidation"*, *"pneumonia"*, *"atelectasis"*, *"pneumothorax"*, *"pleural effusion"*, *"pleural thickening"*, *"fracture"*, *"support devices"*, and *"no finding"*.

Construction Procedure. Given the set of pathology concepts $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, where each c_k corresponds to one of the predefined CheXpert pathology labels, we construct each prototype p_k using the following procedure:

1. **Tokenization:** Each concept name c_k is tokenized using the text encoder's tokenizer (WordPiece Tokenizer trained on CheXpert data) to produce a sequence of token IDs t_k.
2. **Text Encoding:** The tokenized sequence is passed through the text encoder $\psi_T(\cdot)$ to obtain contextualized token embeddings $h_k \in \mathbb{R}^{L_k \times D_T}$, where L_k is the sequence length for concept c_k.
3. **Projection:** The encoded representations are projected into the shared latent space using the text projection function $\phi_T(\cdot)$, yielding $\phi_T(h_k) \in \mathbb{R}^{L_k \times D}$.
4. **Mean Pooling:** To obtain a single prototype vector per concept, we compute the mean across the sequence dimension:

$$p_k = \frac{1}{L_k} \sum_{i=1}^{L_k} \phi_T(h_k)_i \in \mathbb{R}^D. \tag{1}$$

This procedure ensures that each prototype p_k encodes the semantic meaning of its corresponding pathology concept in the shared vision-language embedding space. These prototypes remain fixed during training and serve as semantic anchors for aligning visual and textual features.

A.2. Additional Architectural Details

This subsection expands on the model architecture specifications briefly mentioned in Section 4.2 (Implementation Details) of the main paper, providing complete architectural specifications for all components.

Model Specifications. The visual encoder ψ_I uses DINOv2-ViT-B/14, outputting $N_I = 1370$ patch tokens (corresponding to 518×518 resolution) with dimension $D_I = 768$. The text encoder ψ_T is CXRBert-general with dimension $D_T = 768$. Both projection functions ϕ_I and ϕ_T are 2-layer MLPs with hidden dimension 1536, projecting to shared latent space dimension $D = 768$ using GELU activation and layer normalization.

The SCPR decoder D_ω is a 4-layer transformer decoder with 8 attention heads, feedforward dimension $d_{\text{ff}} = 3072$

and hidden dimension $d_{\text{model}} = 768$, with masking ratio 40% during pre-training. The Multimodal Fusion module is a 2-layer transformer encoder with 8 heads and hidden dimension 768. The report generator uses DistilGPT2 (6 layers, 12 heads, dimension 768) with maximum generation length 100 tokens.

A.3. Training Details

This subsection provides complete training hyperparameters and schedules that extend the brief description in Section 4.2 of the main paper.

Optimization and Schedules. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, and gradient clipping with maximum norm 1.0. Pre-training runs for 20 epochs with 1000-step linear warmup, while fine-tuning runs for 50 epochs with 500-step warmup.

PAPA Hyperparameters. For the optimal transport assignment in Section 3.3, we use Sinkhorn iterations $L = 3$, temperature $\tau = 0.1$, and entropy regularization $\epsilon = 0.05$. The loss balancing coefficients are set to $\lambda = 1.0$, $\lambda_1 = 1.0$, and $\lambda_2 = 0.1$.

Data Preprocessing. Images are resized to 518×518 pixels. For multi-view inputs, we use multiviews dataset from MLRG with the maximum number of view $n = 2$. Text sequences use maximum length 256 tokens for reports and 64 tokens for indications.

Computational Setup. Experiments use 1 NVIDIA RTX 4090 GPU (24GB). Pre-training takes $\sim$12 hours and fine-tuning takes $\sim$18 hours on MIMIC-CXR.

A.4. Sinkhorn-Knopp Algorithm for Prototype Assignment

This subsection provides the complete algorithmic details for the Sinkhorn-Knopp algorithm referenced in Section 3.3 of the main paper (PAPA mechanism).

As described in Section 3.3, we use the Sinkhorn-Knopp algorithm to solve the entropy-regularized optimal transport problem for assigning features to pathology prototypes. Given feature representations $F \in \mathbb{R}^{N \times D}$ (either visual or textual `[CLS]` tokens) and pathology prototypes $P \in \mathbb{R}^{K \times D}$, the algorithm computes a soft assignment matrix $\mathcal{T} \in \mathbb{R}^{N \times K}$ through alternating row and column normalization on the kernel matrix $K = \exp(-C/\epsilon)$, where C encodes feature-prototype similarities and ϵ controls entropy regularization. This ensures balanced assignments that prevent mode collapse while establishing pathology-level correspondences between modalities.

Algorithm 1 Sinkhorn-Knopp algorithm for Prototype Assignment.

Input: Feature $F \in \mathbb{R}^{N \times D}$, prototypes $P \in \mathbb{R}^{K \times D}$, regularization ϵ, temperature τ, iterations L
Compute cost matrix: $C_{ij} = -\dfrac{F_i^\top p_j}{\tau}$
Compute kernel matrix: $K = \exp(-C/\epsilon)$
Initialize scaling vectors: $u = \mathbf{1}_N/N, \quad v = \mathbf{1}_K/K$
for $l = 1$ **to** L **do**
    $u \leftarrow \dfrac{1/N}{Kv}$
    $v \leftarrow \dfrac{1/K}{K^\top u}$
end for
Compute transport plan: $\mathcal{T} = \text{diag}(u)\, K\, \text{diag}(v)$
Output: Transport plan $\mathcal{T}$

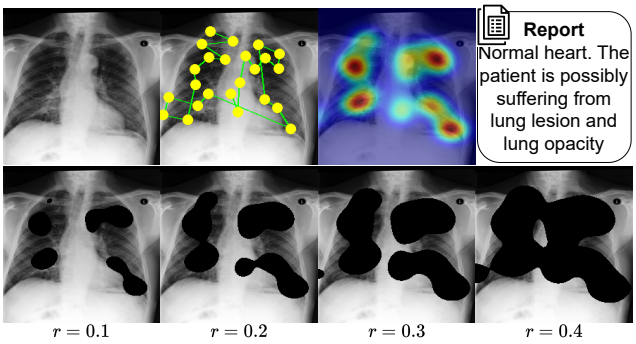

Figure S-II. **Anatomical-specific masking methodology using radiologist eye gaze annotations.** *Top row:* From left to right: original chest X-ray, radiologist eye gaze fixation points (yellow dots with green connections), eye gaze heatmap identifying anatomical regions (red indicates high attention), and the corresponding radiology report indicating lung lesion and lung opacity findings. *Bottom row:* Progressive masking at increasing ratios ($r = 0.1, 0.2, 0.3, 0.4$).

A.5. Mask Creation Methodology

To evaluate the visual grounding capability of our model and baseline methods, we conduct controlled masking experiments on chest X-ray images. The masking procedure simulates occlusion of anatomical regions. We employ different masking strategies during training and inference evaluation.

Training: Random Masking. During pre-training with SCPR (Section 3.2), we use random patch masking to encourage spatially consistent representations. To apply masking at the patch level, we divide the input image into $N_p = 1370$ patches, which do not form a square grid. Given an image $I \in \mathbb{R}^{C \times H \times W}$, we first estimate the spatial patch layout by setting the number of columns to $N_{\text{cols}} =$

$\lfloor \sqrt{N_p} \rfloor$, and the number of rows to $N_{\text{rows}} = \left\lceil \frac{N_p}{N_{\text{cols}}} \right\rceil$. For an image of size 518×518, this produces an uneven grid. Each patch therefore has spatial dimensions $p_h = \left\lfloor \frac{H}{N_{\text{rows}}} \right\rfloor$ and $p_w = \left\lfloor \frac{W}{N_{\text{cols}}} \right\rfloor$

A binary mask of length N_p is sampled for each image and reshaped into a $(N_{\text{rows}} \times N_{\text{cols}})$ grid, with zero-padding if necessary. For each masked patch location, the corresponding spatial region in the image is set to zero.

For each image in a batch of size B, we randomly select $\lfloor r \cdot N_p \rfloor$ patches to mask, where $r = 0.4$ is the masking ratio. The selection is performed independently for each sample using a random permutation of patch indices:

$$\mathcal{M}_b = \{\pi_b(i) : i = 1, \ldots, \lfloor r \cdot N_p \rfloor\}, \tag{2}$$

where π_b denotes a random permutation for the b-th sample, and $\mathcal{M}_b$ is the set of masked patch indices. This random masking prevents the model from exploiting spatial biases and ensures robust feature learning across all image regions.

For each selected patch at grid position (i, j), we compute the spatial coordinates in the image:

$$h_0 = i \cdot p_h, \quad h_1 = \min((i+1) \cdot p_h, H), \tag{3}$$

$$w_0 = j \cdot p_w, \quad w_1 = \min((j+1) \cdot p_w, W), \tag{4}$$

and set the corresponding pixel values to zero: $I[:, h_0{:}h_1, w_0{:}w_1] = 0$. This creates a binary patch mask $\mathbf{M} \in \{0,1\}^{N_p}$ where $\mathbf{M}_{ij} = 1$ indicates a masked patch.

This dual masking strategy allows us to both train robust spatial representations and systematically evaluate whether models genuinely rely on visual pathology information (showing degradation under anatomical masking) or exploit language priors (maintaining performance despite visual perturbations).

Inference: Anatomical-Specific Masking. For evaluation experiments (reported in Section 4.5 and Table S-I), we employ targeted masking to assess whether models genuinely rely on visual information. We utilize the EYEGAZE dataset, which provides radiologist gaze-based annotations indicating anatomical regions for 1,083 CXR images.

We evaluate under two conditions: (1) *Anatomical Masking* where we use the available anatomical heatmaps from EYEGAZE dataset . We aggregate and resize all heatmaps for each image, forming a combined heatmap that highlights anatomical regions. We then derive a binary mask using an adaptive threshold that selects the top-activated pixels to match a target masking ratio $r \in \{0.1, 0.2, 0.3, 0.4\}$. (2) *Non-Anatomical Masking* where we insert random black rectangles while explicitly preventing any overlap with anatomical regions. Rectangles of varying sizes are repeatedly sampled until the target masked area is reached or the sampling limit is met, ensuring that only visually uninformative, non-anatomical regions are masked.

B. Experimental Results

This section provides comprehensive experimental results that extend the main paper. We present detailed masking experiments and per-pathology clinical efficacy analyses.

B.1. Metrics

We evaluate generated radiology reports using a combination of linguistic and clinical metrics. Linguistic metrics assess surface-level text quality and fluency, while clinical metrics evaluate diagnostic accuracy and correctness of medical content. Given a generated report p and a ground-truth report g, the metrics are defined as follows.

Linguistic Metrics. BLEU-n (B-n) measures n-gram precision with brevity penalty:

$$\begin{aligned} \text{BLEU-n} &= \text{BP} \cdot \exp\left(\sum_{k=1}^{n} w_k \log p_k\right), \\ p_k &= \frac{\text{\# of matched } k\text{-grams}}{\text{\# of } k\text{-grams in } p}, \\ \text{BP} &= \begin{cases} 1, & \text{if } |p| > |g| \\ \exp\left(1 - |g|/|p|\right), & \text{otherwise} \end{cases}, \\ w_k &= \frac{1}{n}. \end{aligned} \tag{5}$$

ROUGE-L (R-L) evaluates the longest common subsequence (LCS) between p and g:

$$\begin{aligned} P_L &= \frac{LCS(p,g)}{|p|}, \quad R_L = \frac{LCS(p,g)}{|g|}, \\ F1_L &= \frac{(1+\beta^2) P_L R_L}{R_L + \beta^2 P_L}, \\ \beta &= \frac{|g|}{|p|}. \end{aligned} \tag{6}$$

METEOR (MTR) evaluates the harmonic mean of precision and recall at the unigram level, with recall weighted more heavily and incorporating synonyms and stemming:

$$\begin{aligned} \text{MTR} &= F_{\text{mean}} \cdot (1 - \text{Penalty}), \\ F_{\text{mean}} &= \frac{10\,P\,R}{R + 9\,P}, \\ \text{Penalty} &= 0.5 \left(\frac{\text{chunks}}{\text{matches}}\right)^3, \end{aligned} \tag{7}$$

where P and R are the unigram precision and recall, matches is the number of aligned unigrams, and chunks is the number of contiguous matched sequences.

Table S-I. Comparison of **LLaVARad**, **MLRG**, and **SCOPE** under different masking transformations and ratios. Evaluation metrics include BLEU-4 (B-4), ROUGE-L (R-L), and F1-RadGraph (RG). Values in parentheses indicate percentage drop from baseline (0.0 ratio). For *Non-Anatomical Masking*, smaller drops indicate robustness to irrelevant occlusions. For *Anatomical Masking*, larger drops demonstrate genuine visual grounding. The last row is the number of parameters and inference time of each models.

Transformation	Ratio	LLaVARad			MLRG			SCOPE (Ours)		
		B-4	**R-L**	**RG**	**B-4**	**R-L**	**RG**	**B-4**	**R-L**	**RG**
Non-Anatomical Masking	0.4	0.152 (-10.1)	0.301 (-8.8)	0.311 (-11.1)	0.217 (-9.6)	0.380 (-5.2)	0.361 (-5.2)	0.237 (**-2.9**)	0.402 (**-2.4**)	0.402 (**-2.9**)
	0.3	0.156 (-7.7)	0.312 (-5.5)	0.335 (-4.3)	0.218 (-9.2)	0.378 (-5.7)	0.363 (-4.7)	0.239 (**-2.0**)	0.407 (**-1.2**)	0.404 (**-2.4**)
	0.2	0.160 (-5.3)	0.320 (-3.0)	0.342 (-2.3)	0.213 (-11.2)	0.367 (-8.5)	0.354 (-7.1)	0.243 (**-0.4**)	0.407 (**-1.2**)	0.409 (**-1.2**)
	0.1	0.165 (-2.4)	0.327 (-0.9)	0.347 (-0.9)	0.231 (-3.7)	0.389 (-3.0)	0.374 (-1.8)	0.243 (**-0.4**)	0.409 (**-0.7**)	0.411 (**-0.7**)
	0.0	0.169	0.330	0.350	0.240	0.401	0.381	0.244	0.412	0.414
Anatomical Mask	0.4	0.150 (-11.2)	0.298 (-9.7)	0.329 (-6.0)	0.171 (**-28.7**)	0.349 (**-13.0**)	0.351 (-7.9)	0.192 (-21.3)	0.363 (-11.9)	0.343 (**-17.1**)
	0.3	0.153 (-9.5)	0.304 (-7.9)	0.325 (-7.1)	0.182 (**-24.2**)	0.348 (**-13.2**)	0.346 (-9.2)	0.213 (-12.7)	0.376 (-8.7)	0.357 (**-13.8**)
	0.2	0.157 (-7.1)	0.310 (-6.1)	0.335 (-4.3)	0.193 (**-19.6**)	0.360 (**-10.2**)	0.349 (-8.4)	0.215 (-11.9)	0.381 (-7.5)	0.363 (**-12.3**)
	0.1	0.163 (-3.6)	0.320 (-3.0)	0.341 (-2.6)	0.207 (**-13.8**)	0.371 (**-7.5**)	0.350 (-8.1)	0.227 (-7.0)	0.384 (-6.8)	0.371 (**-10.4**)
	0.0	0.169	0.330	0.350	0.240	0.401	0.381	0.244	0.412	0.414
Params / Inference Time		∼7B / 1.71s			296M / 0.181s			306M / 0.191s		

Clinical Metrics. For clinical evaluation, we employ CheXbert to label reports with 14 predefined clinical findings. For each class i, the per-class precision, recall, and F1-score are defined as:

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad R_i = \frac{TP_i}{TP_i + FN_i}, \quad F1_i = \frac{2P_iR_i}{P_i + R_i}. \tag{8}$$

Micro-averaged scores aggregate counts before computing metrics, while macro-averaged scores compute per-class metrics first then average:

$$\begin{aligned} P_{\text{micro}} &= \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}, \quad P_{\text{macro}} = \frac{1}{C}\sum_i P_i, \\ R_{\text{micro}} &= \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}, \quad R_{\text{macro}} = \frac{1}{C}\sum_i R_i, \\ F1_{\text{micro}} &= \frac{2P_{\text{micro}}R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}, \quad F1_{\text{macro}} = \frac{1}{C}\sum_i F1_i, \end{aligned} \tag{9}$$

where C is the number of clinical classes.

Finally, RadGraph F1 (RG) evaluates the correctness of extracted entities and relations:

$$RG = \frac{2 \cdot TP_{\text{ent/rel}}}{2 \cdot TP_{\text{ent/rel}} + FP_{\text{ent/rel}} + FN_{\text{ent/rel}}}. \tag{10}$$

B.2. Visual Masking Experiments

This subsection extends the visual masking experiments presented in Section 4.5 of the main paper and provides comprehensive quantitative results across all metrics.

Table S-I extends Table 3 from the main paper (Section 4.5) by including all metrics (BLEU-4, ROUGE-L, and F1-RadGraph) for comprehensive comparison. We design controlled masking experiments to assess whether SCOPE depends on visual pathology cues, unlike baseline methods (MLRG, LLaVARad) that rely on language priors. If SCOPE's predictions are visually grounded, performance should noticeably decline when key anatomical regions are occluded, whereas text-driven baselines should remain relatively unaffected.

Non-Anatomical Masking Robustness. When non-diagnostic regions are masked, SCOPE achieves the smallest performance drops across all metrics and ratios (bold in Table S-I). At the highest masking ratio (0.4), SCOPE shows minimal degradation (B-4: -2.9%, R-L: -2.4%, RG: -2.9%) compared to MLRG (-5.2% to -9.6%) and LLaVARad (-8.8% to -11.1%), indicating that SCOPE does not rely on spurious correlations from non-pathological regions.

Anatomical Masking Sensitivity. Conversely, when diagnostically critical regions are occluded, SCOPE demonstrates genuine visual grounding through substantial performance drops. For the clinically most important F1-RadGraph metric, SCOPE achieves the largest degradation across all ratios (e.g., -17.1% at 0.4 masking), significantly exceeding MLRG (-7.9%) and LLaVARad (-6.0%). This 2-3× greater sensitivity to anatomical masking demonstrates strong dependency on visual pathology information, whereas LLaVARad's minimal sensitivity indicates heavy reliance on language priors.

Optimal Visual Grounding. SCOPE exhibits ideal behavior: robust to irrelevant occlusions yet sensitive to pathological region masking. SCOPE achieves these gains while maintaining competitive efficiency (306M parameters, 0.191s inference) comparable to MLRG (296M, 0.181s), whereas LLaVARad requires substantially more resources (∼7B, 1.71s) yet shows weaker visual grounding.

B.3. Per-Pathology Clinical Efficacy Results

Table S-II provides detailed per-pathology clinical efficacy results that complement Table 2 in the main paper (Section 4.3), showing precision, recall, and F1-score breakdowns for each of the 14 clinical findings on MIMIC-CXR dataset.

Visual Grounding Benefits Across Pathology Frequencies. Table S-II shows per-pathology clinical efficacy met-

Table S-II. Comparison of SEI , MLRG, and our method in terms of clinical accuracy on the MIMIC-CXR dataset, where P, R, and F1 denote Precision, Recall, and F1-score, respectively. Win/Loss columns show percentage improvements (+) or degradations (-) of our method vs. MLRG, computed as $\Delta\% = (v_{\text{ours}} - v_{\text{MLRG}})/v_{\text{MLRG}} \times 100\%$, where v_{ours} and v_{MLRG} represent the metric values (P, R, or F1) for our method and MLRG, respectively.

Finding	Freq. (%)	SEI			MLRG			Ours			Win/Loss (%)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Cardiomegaly	14.8	0.599	0.633	0.616	0.629	0.570	0.598	0.684	0.651	0.668	+8.7	+14.2	+11.7
Lung Opacity	13.8	0.519	0.170	0.256	0.594	0.317	0.413	0.620	0.345	0.443	+4.4	+8.8	+7.3
Support Devices	12.8	0.763	0.708	0.734	0.768	0.788	0.778	0.816	0.809	0.813	+6.3	+2.7	+4.5
Pleural Effusion	12.4	0.683	0.697	0.690	0.716	0.641	0.676	0.763	0.636	0.694	+6.6	-0.8	+2.7
Atelectasis	10.9	0.469	0.395	0.429	0.499	0.475	0.487	0.522	0.457	0.487	+4.6	-3.8	+0.0
Enlarged Cardiomediastinum	10.0	0.373	0.208	0.267	0.370	0.353	0.361	0.459	0.465	0.462	+24.1	+31.7	+28.0
Edema	8.3	0.526	0.361	0.428	0.516	0.448	0.480	0.605	0.479	0.535	+17.2	+6.9	+11.5
Pneumonia	4.4	0.174	0.065	0.095	0.316	0.235	0.270	0.364	0.246	0.293	+15.2	+4.7	+8.5
Consolidation	3.3	0.218	0.194	0.205	0.259	0.150	0.190	0.317	0.159	0.212	+22.4	+6.0	+11.6
Lung Lesion	2.5	0.462	0.021	0.041	0.429	0.046	0.082	0.638	0.114	0.194	+48.7	+147.8	+136.6
No Finding	2.4	0.161	0.597	0.253	0.233	0.629	0.340	0.249	0.685	0.365	+6.9	+8.9	+7.4
Fracture	1.8	0.000	0.000	0.000	0.174	0.021	0.037	0.361	0.067	0.113	+107.5	+219.0	+205.4
Pleural Other	1.6	0.167	0.022	0.039	0.231	0.054	0.087	0.338	0.119	0.176	+46.3	+120.4	+102.3
Pneumothorax	1.0	0.174	0.039	0.064	0.426	0.230	0.299	0.533	0.160	0.246	+25.1	-30.4	-17.7
micro avg	-	0.523	0.410	0.460	0.549	0.468	0.505	0.597	0.500	0.545	+8.7	+6.8	+7.9
macro avg	-	0.378	0.294	0.294	0.440	0.354	0.364	0.519	0.385	0.407	+18.0	+8.8	+11.8

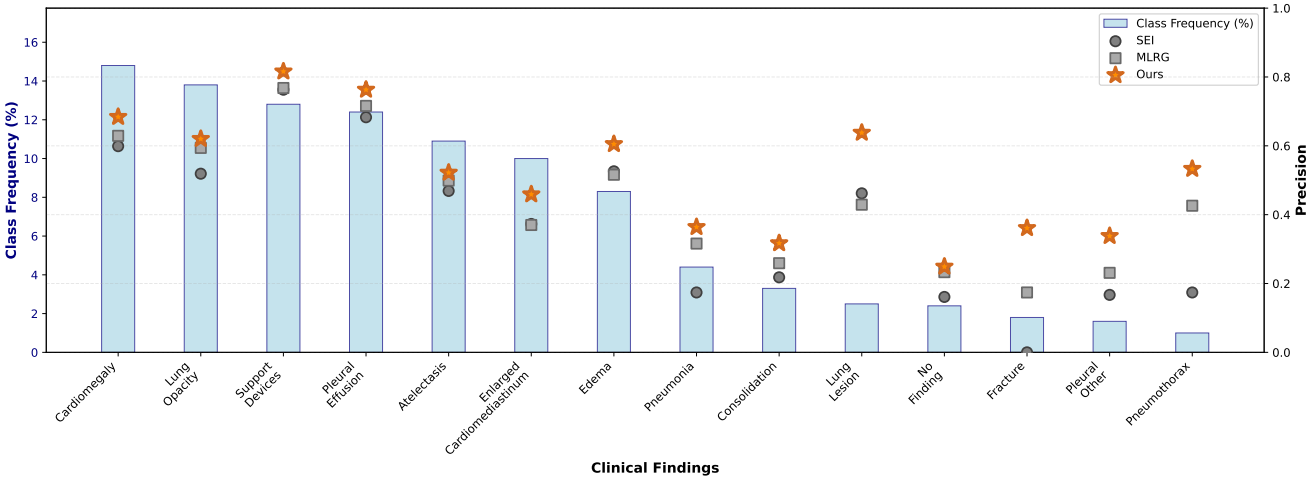

Figure S-III. **Per-pathology precision versus class frequency distribution on MIMIC-CXR.** Blue bars show class frequency (left y-axis) for 14 clinical findings sorted in descending order, revealing a pronounced long-tail distribution ranging from Cardiomegaly (14.8%) to Pneumothorax (1.0%). Overlaid scatter points show precision scores (right y-axis) for SEI (gray circles), MLRG (gray squares), and SCOPE (orange stars). SCOPE achieves consistently superior precision across all pathologies, with particularly dramatic improvements on rare conditions in the distribution tail (Fracture, Lung Lesion, Pleural Other, Pneumothorax), where language-prior-driven baselines cannot exploit memorized co-occurrence patterns. The widening performance gap at lower frequencies validates that SCOPE's visual grounding becomes comparatively more beneficial when linguistic shortcuts are unavailable.

rics on MIMIC-CXR. SCOPE achieves consistent improvements across most pathologies compared to baselines. The macro-average F1 improvement (+11.8%) exceeds the micro-average gain (+7.9%), indicating stronger performance on less frequent findings.

Examining individual pathologies, SCOPE shows notable gains on rare conditions. For instance, Fracture (1.8% frequency) improves from 0.037 F1 (MLRG) to 0.113 F1 (+205.4%), Pleural Other (1.6%) from 0.087 to 0.176 (+102.3%), and Lung Lesion (2.5%) from 0.082 to 0.194 (+136.6%). On more frequent pathologies like Cardiomegaly (14.8%) and Support Devices (12.8%), improvements are more modest at +11.7% and +4.5% respectively. Figure S-III visualizes this pattern, where precision gains (orange stars vs. gray markers) are particularly pronounced for findings on the right tail of the frequency distribution.

This observation is consistent with the visual grounding hypothesis presented in the main paper. Since rare patholo-

gies have limited co-occurrence patterns in training corpora, language-prior-driven baselines cannot rely on memorized textual associations. SCOPE's visual grounding through PAPA and SCPR becomes comparatively more beneficial in these cases where linguistic shortcuts are unavailable.