CheXmix: Unified Generative Pretraining for Vision Language Models in Medical Imaging

Supplementary Material

Contents

A. CheXmix Training Hyperparameters

Configuration	CheXmix (S1)	CheXmix (S2)
LLM Init.	RadPhi-2	from Stage 1
Image Resolution	512^2	512^2
Masking Percentage	-	50%
LLM max sequence length	1300	1300
Optimizer	AdamW	
Optimizer hyperparameter	$\beta_1 = 0.9$, $\beta_2 = 0.98$, *eps* $= 1\text{e} - 6$	
Peak learning rate	1e-5	1e-5
Learning rate schedule	cosine decay	
Weight decay	0.1	0.1
Gradient clip	1.0	1.0
Training steps	703,671	513,993
Global batch size	8	8
Gradient Acc.	4	4
Numerical precision	bfloat16	bfloat16

Table 1. Pretraining hyperparameters for CheXmix S1 and S2 models.

B. Extended Methods

B.1. Model Pretraining

Additional Staged Pretraining Explanation We provide a simplified explanation below to clarify our training objectives and naming conventions. Both stages use a next-token prediction loss, but differ in whether input tokens are masked. We provide detailed hyperparameters for pretraining in Table 1. Our models are pretrained on both chest X-ray image tokens and radiology report text tokens using a two-stage approach, detailed below:

1. **Stage 1: Standard Autoregressive Pretraining:** This is the conventional autoregressive language modeling objective. Given an input sequence $(t_1, t_2, \ldots, t_N)$, the model predicts the next token at each position, and the cross-entropy loss is summed over the entire sequence.
2. **Stage 2: Masked Image-Language Pretraining:** In Stage 2, we introduce an autoregressive masked-token prediction loss that combines ideas from autoregressive modeling and masked autoencoders. Similar to Stage 1, the model predicts the next token at each position; however, we randomly replace 50% of the input image and text tokens with a special `[MASK]` token, and we compute the loss *only* for output tokens that immediately follow a masked input token (Figure 2b). In effect, the model must reconstruct information missing from its input by leveraging corrupted context.

   For example, if the input sequence is (t_1, t_2, t_3) and we randomly mask position 2, we obtain $(t_1, \texttt{[MASK]}, t_3)$. The model then predicts (p_2, p_3, p_4). We ignore the losses for p_2 (predicted after t_1) and p_4 (predicted after t_3), and compute the loss only for p_3, which is predicted after the masked token and is compared against the ground-truth t_3. This design encourages the model to accurately predict tokens following masked inputs, and our evaluations indicate strong representation quality and improved robustness to masked input under this strategy.

B.2. Evaluation Methods

We present details of our evaluation pipeline. In general, we conduct a rigorous analysis of representational quality through both discriminative and generative tasks. Our evaluation suite includes CheXpert embedding findings classification, image inpainting, radiology report generation, multimodal retrieval, test-time augmentation for report generation, and several ablation studies. We first assess the discriminative capability of CheXmix's embeddings by evaluating pretrained representations on the CheXpert dataset and comparing them to relevant general-domain and medical-specific baselines. Next, we evaluate CheXmix's generative capabilities through image inpainting and radiology report generation, using a test set composed of images and reports from five datasets (Section 3.1; Main Paper).

For both classification and generation tasks, we examine model performance across multiple masking percentages (20%, 40%, 60%, and 80%) to highlight CheXmix's fine-grained representational capacity. The rationale for masking during evaluation is to provide a general assessment of each model's representational robustness to partial or occluded inputs; in real-world chest radiographs, regions of interest may be obscured by medical devices (e.g., pacemakers, ECG leads) [8], imaging artifacts [11], or overlapping anatomy [10], and a robust model should leverage global context to make accurate predictions despite missing information. We further demonstrate improvements in report quality using CheXmix's test-time augmentation strategy, which does not require additional training, and we evaluate the impact of causal masking versus bidirectional attention during pretraining, as well as the effect of different masking ratios.

CheXpert Findings Classification: We evaluate pretrained representations on the CheXpert dataset using a multi-head masked linear probe classification task over 14 findings: *Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices, and No Finding*. The pretrained embedding dimensions are as follows: Chameleon (4096), HealthGPT (1024), MAE (384), M3AE (768), CheXagent (2560), CheXmix (Stage 1 and 2, 2560). For CheXmix, we tokenize images using the VQ-GAN tokenizer developed with Chameleon [1]. We first process the embeddings for the images on the CheXpert dataset and take the mean of the embeddings across all image patches to get a single embedding vector for the image. Averaging token embeddings has been shown to result in better performance than probing at other token positions [16].

In generative models, unlike discriminative models, the information relevant for classification is not necessarily concentrated at the final layer. Therefore, we extract embeddings from every layer of each model. Specifically, we consider Chameleon (32 layers), HealthGPT (24 layers), MAE (13 layers), M3AE (13 layers), CheXagent (26 layers), and CheXmix (32 layers). For each model, we select the layer that achieves the highest AUROC on the validation set and report the corresponding AUROC and AUPRC metrics on the test set. The middle layers yield the best embeddings for generatively pretrained models (Chameleon, CheXmix), whereas the final layers perform best for vision encoder models (M3AE, CheXagent) (Table 2).

For each model, we train 14 linear probes with bias terms, optimized using AdamW (without weight decay) for 100 epochs with batch size 8 and gradient accumulation of 8. Training is conducted with an initial learning rate of 1×10^{-5}, cosine learning rate decay, and without mixed-precision. To examine robustness under different levels of supervision, we apply masking ratios of 20%, 40%, 60%, and 80% during training. For the generatively pretrained models (Chameleon, MAE, M3AE, CheXmix), masking is applied at the token level, whereas for CheXagent and HealthGPT, which inputs image patches directly into the transformer, masking is applied at the image level. CheXpert labels are provided as -1 (uncertain/missing), 0, 1, or empty; we treat empty cells as -1 and mask the loss for labels equal to -1. We process the CheXpert training data and generate train/validation/test splits of 22,342 / 234 / 667, respectively. Performance is

measured using AUROC and AUPRC and all reported metrics metrics are averaged over three random seeds. The goal of this task is to isolate and evaluate the quality of pretrained representations.

Model	**Masking Percentage (%)**				
	0%	**20%**	**40%**	**60%**	**80%**
Chameleon	4	4	7	18	1
HealthGPT	23	17	23	18	17
MAE	11	9	8	9	11
M3AE	Trans. 10	Trans. 8	Trans. 8	Trans. 9	Trans. 10
CheXmix (S1)	8	10	9	8	18
CheXmix (S1 + S2)	8	8	8	8	8
CheXmix (S1; B)	14	15	12	12	7
CheXmix (S2; B)	10	3	5	5	32
CheXagent	Encoder 23	Encoder 16	Encoder 10	Encoder 10	Encoder 8

Table 2. Best-Performing Layers Across Masking Percentages. Models vary in their number of layers (e.g., CheXmix has 33). For generatively pretrained models (Chameleon, CheXmix), intermediate layers produce the strongest embeddings, whereas for vision encoder models (M3AE, CheXagent), the final layers perform best.

Image Inpainting: We evaluate inpainting performance by reconstructing masked regions of images and assessing quality using similarity metrics including PSNR, MS-SSIM, and FID-Inception, implemented via the `torchmetrics` library. For evaluation, we randomly sample 5,000 images from the validation split of our pretraining dataset, which consists of five different datasets, and tokenize them with the VQ-GAN tokenizer [1]. We experiment with masking ratios ranging from 10–90%. For the CheXmix (Stage 1 and 2) models, the designated mask token is 58,560. At each masking ratio, we randomly generate indices to mask within the token sequence and feed the partially masked sequence through the model, predicting replacements by selecting the token with the highest probability score. We then decode the predicted tokens back into image space. As a baseline, we also measure reconstruction quality by tokenizing an image, applying masking, and then directly decoding the tokens back into image space using the VQ-GAN decoder without generative modeling. The goal of this task is to measure the model's ability to reconstruct high-fidelity visual details from incomplete observations.

Radiology Report Generation: We evaluate radiology report generation by providing images with varying levels of masked input and generating the corresponding reports, focusing on both the findings and impression sections. Performance is assessed using domain-specific metrics including GREEN [9], CheXbert [12], RadGraph-F1 [6], and BERTScore [15]. For evaluation, we randomly sample 1000 image–text pairs from the validation split of our pretraining dataset, which consists of five different datasets, and tokenize the images using a VQ-GAN tokenizer. We experiment with masking ratios ranging from 10–90%. At each ratio, we randomly generate indices to mask within the token sequence and feed the partially masked sequence and T_S (text-start token) through the model to generate the associated report. To ensure fair comparison across samples, we set the maximum token limit for generated reports equal to the length of the original input report. As a baseline, we also evaluate the Chameleon model, which is masked at the token level. For Chameleon, we provide the prompt: `Generate a findings and impression section for this chest X-ray image. Include the 'findings' and 'impressions' tag in the report. Do not`

`list the findings and impressions separately; instead, present them in one continuous section.` For CheXagent, we provide the prompt: `Generate the findings and impression section.`

Radiology Report Generation (Test-Time Augmentation): To illustrate a practical use case of masked learning and better motivate CheXmix (S2) pretraining, we process over 1,000 generated reports using a Test-Time Augmentation (TTA) strategy. Specifically, we introduce a disjoint masking protocol that generates multiple, distinct masked versions of each image, allowing the model to generate radiology reports from these masked image tokens.

Let $\mathcal{I} = \{1, \ldots, N\}$ be the set of all $N = 1024$ image token indices. We partition $\mathcal{I}$ into $K = 5$ mutually disjoint subsets $\mathcal{S}_1, \ldots, \mathcal{S}_K$, such that every token belongs to exactly one subset:

$$\bigcup_{k=1}^{K} \mathcal{S}_k = \mathcal{I} \quad \text{and} \quad \mathcal{S}_k \cap \mathcal{S}_j = \emptyset, \quad \forall k \neq j. \tag{1}$$

This partition ensures that the subsets $\mathcal{S}_k$ are unique and non-overlapping. Using this partition, we define the set of *visible* tokens, denoted as $\mathcal{V}_k$, for the k-th input variation under two distinct settings:

1. **20% Masking (Disjoint Masks):** In this setting, we mask the tokens in $\mathcal{S}_k$ while keeping the remainder visible. Formally, the visible set is defined as the complement:

$$\mathcal{V}_k = \mathcal{I} \setminus \mathcal{S}_k. \tag{2}$$

   Across the K variations, the masked region shifts such that a different 20% of the image is hidden in each pass, allowing the model to generate radiology reports from complementary subsets of visual information.
2. **80% Masking (Disjoint Views):** In this setting, we keep *only* the tokens in $\mathcal{S}_k$ visible, masking the remaining 80%. The visible set is:

$$\mathcal{V}_k = \mathcal{S}_k. \tag{3}$$

   Here, the unmasked regions are disjoint across the K variations. This forces the model to attend to distinct visual information in each pass for generating radiology reports.

Gemini 2.5 Pro [4] (`gemini-2.5-pro`) then consolidates the unique characteristics from the five radiology reports, generated from the 20% and 80% masked inputs, into a single synthesized report. We evaluate the performance of this TTA strategy using the GREEN and CheXbert metrics, producing reports as described in Section B.2. The primary intuition behind TTA is that sampling the model multiple times with different masked and unmasked image inputs captures variations in its predictions, thereby probing the model's epistemic uncertainty. An example of the Gemini prompt used in this process is provided in Section E.

Multimodal Retrieval: We evaluate image–text (chest X-ray $\rightarrow$ radiology report) and text–image (radiology report $\rightarrow$ chest X-ray) retrieval performance for CheXmix (S1), CheXmix (S1+S2), and CheXagent (SigLIP). Retrieval is assessed using Top-8 and Top-16 accuracy across chunked pool sizes of 32, 64, and 128 over 2,048 test samples. For each image–text pair, we compute cosine similarity between the corresponding embeddings and rank them accordingly. Chest X-rays are 512×512 images tokenized into 1,024 discrete tokens and processed through the unified transformer decoder for CheXmix, and patch-wise (32×32) through the vision transformer for SigLIP. Radiology reports, composed of the Findings and Impression sections, are encoded using the CheXmix unified transformer decoder or the SigLIP text encoder. The resulting embeddings are 2,056-dimensional for both image and text in CheXmix, and 1,024-dimensional for both

modalities in SigLIP. For all models, we average token embeddings to produce a single vector representation per modality for retrieval.

Pretraining Ablation (Masking Ratio): We evaluate the effect of masking ratio during CheXmix (S2) pretraining by training models with four ratios: 25%, 50%, 75%, and 90%. These values are motivated by medical-domain literature [14], which shows that moderate masking ratios (e.g., 40%) can yield strong representations for chest X-rays, and the seminal masked autoencoder work [5], which recommends higher ratios such as 75% or 90% to reduce spatial redundancy. We pretrain all models for 100K steps and then assess their downstream performance on the CheXpert embedding–based findings classification task. For each masking ratio, we compute AUROC and AUPRC by selecting the best-performing layer among the 32 layers of the transformer decoder using the validation set, and we report the corresponding performance on the test set.

Pretraining Ablation (Causal Mask): We evaluate pretraining with and without applying a causal mask to the image tokens at both CheXmix S1 and S2 pretraining stages. Recent unified generative models in the general domain, such as Show-O [13] and Transfusion [17], report strong performance when images are given full bidirectional attention while text tokens remain causally masked. This design reflects the intuition behind vision transformers, where bidirectional attention is appropriate because the causal ordering of image patches is not semantically meaningful. In contrast, other models such as Chameleon [1] maintain causal masking for both modalities.

In our approach, we enable full bidirectional attention over the 1,024 image tokens while keeping the text tokens strictly causal. We construct the attention mask by first initializing a standard causal mask $A \in \mathbb{R}^{N \times N}$ for the entire sequence of $N = N_{\text{img}} + N_{\text{text}}$ tokens. To convert the image portion to full attention, we overwrite the corresponding block of the attention matrix with zeros. Specifically, for image token indices $i, j \in \{1, \ldots, N_{\text{img}}\}$, we update the mask such that $A_{ij} = 0$, ensuring bidirectional attention among image tokens while preserving causal masking for all positions involving text tokens. For detailed hyperparameters and training configurations, please refer to Table 3.

NIH ChestX-ray14 (External Evaluation): We evaluate pretrained representations on the NIH ChestX-ray14 dataset using a multi-head masked linear probe classification task across 14 findings: atelectasis, cardiomegaly, pleural effusion, infiltration, lung mass, lung nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. We preprocess the dataset and create train/validation/test splits of 19,621 / 1,121 / 2,242, respectively. The embedding dimensions of the pretrained models are as follows: Chameleon (4096), HealthGPT (1024), M3AE (768), CheXagent (2560), and CheXmix (S1 + S2; 2560). We train linear probes following the "CheXpert Findings Classification" protocol (see B.2), without evaluating across masking percentages. Performance is reported using AUROC, averaged over three random seeds. The best-performing layers used for evaluation are: Chameleon (layer 16), SigLIP (encoder_23_mean), M3AE (final layer), CheXmix (S1 + S2; layer 22), and HealthGPT (encoder_23_mean).

ReXgradient-160K (External Evaluation): Similar to the "Radiology Report Generation" protocol (see B.2), we evaluate radiology report generation by providing images (no masking) and generating the corresponding reports, focusing on both the findings and impression section. We select 500 random samples from the test set of ReXgradient-160K, specifically chest x-ray and radiology report samples. We evaluate these metrics over three seeds.

Configuration	CheXmix (S1; B)	CheXmix (S2; B)
LLM Init.	RadPhi-2	from Stage 1
Image Resolution	512^2	512^2
Image Attention	Bidirectional (B)	
Text Attention	Causal Mask (CM)	
Masking Percentage	-	50%
LLM max sequence length	1300	1300
Optimizer	AdamW	
Optimizer hyperparameter	$\beta_1 = 0.9, \beta_2 = 0.98, eps = 1\text{e}-6$	
Peak learning rate	1e-5	1e-5
Learning rate schedule	cosine decay	
Weight decay	0.1	0.1
Gradient clip	1.0	1.0
Training steps	683,000	1,353,000
Num. of GPUs (A100)	8	4
Global batch size	16	8
Gradient Acc.	4	4
Numerical precision	bfloat16	bfloat16

Table 3. Pretraining hyperparameters for CheXmix S1 and S2 models with bidirectional attention for the image tokens.

B.3. Evaluation Metrics

CheXpert Embedding Findings Classification: We evaluate the discriminative capability of CheXmix and baseline embeddings on the CheXpert findings classification task using AUROC and AUPRC. For each model, we select the probe (layer) that achieves the highest AUROC on the validation set and report its corresponding performance on the test set.

1. **Area under the receiver operating characteristic curve (AUROC):** AUROC evaluates a binary classification model's ability to differentiate between positive and negative cases across all possible decision thresholds. It is computed as the area under the curve obtained by plotting the True Positive Rate (TPR, or sensitivity) against the False Positive Rate (FPR, equal to 1–specificity). An AUROC of 0.5 indicates performance equivalent to random chance, while a value of 1.0 denotes perfect discrimination.
2. **Area under the precision-recall curve (AUPRC):** AUPRC measures a model's ability to correctly identify positive cases across different decision thresholds, with particular emphasis on performance when the dataset is imbalanced. It is computed as the area under the curve obtained by plotting Precision (positive predictive value) against Recall (sensitivity). AUPRC emphasizes the model's ability to detect the positive class, making it especially informative when positive cases are rare, as is often the case for many abnormalities. A higher AUPRC indicates that the model maintains strong precision even at high recall levels, demonstrating its capacity to correctly identify true positives while minimizing false positives.

Image Inpainting: We evaluate the inpainting performance of CheXmix and baseline models on PSNR, MS-SSIM, and FID.

1. **Peak Signal-to-Noise Ratio (PSNR):** PSNR quantifies pixel-level reconstruction fidelity by measuring the ratio between the square of the maximum possible pixel value and the Mean Squared Error (MSE) between the ground truth and inpainted image, expressed in decibels (dB). A higher PSNR indicates that the generated image is numerically closer to the original in terms of pixel intensity values.
2. **Multi-Scale Structural Similarity Index Measure (MS-SSIM):** MS-SSIM evaluates the perceptual

quality of the reconstruction by analyzing structural similarity across multiple scales and resolutions, capturing both global patterns and fine-grained local details. Ranging from 0 to 1, a higher MS-SSIM indicates that the model has preserved structural integrity, edge definition, and anatomical patterns, which are relevant for chest X-ray interpretation.

3. **Fréchet Inception Distance (FID):** FID assesses the perceptual realism and diversity of generated images by measuring the distance between feature distributions of real and inpainted images in the embedding space of a pre-trained deep neural network (Inception-v3). Unlike PSNR and MS-SSIM, which rely on direct pixel-level comparisons with ground truth, FID evaluates embedding-level similarity by assessing whether the generated distribution matches the statistical and semantic properties of real data. A lower FID score indicates that the inpainted regions exhibit visual features consistent with the original chest X-rays, suggesting high perceptual quality.

Radiology Report Generation: We compute GREEN, CheXbert, RadGraph, and BERTScore to evaluate our generate radiology report across CheXmix and baselines.

1. **Generative Radiology Report Evaluation and Error Notation (GREEN) [9]:** GREEN is a clinically aligned metric that evaluates radiology report quality by identifying and explaining clinically significant errors. Unlike standard metrics such as BLEU or ROUGE, it leverages large language models to detect discrepancies and provides a quantitative score. A higher GREEN score indicates a report that is accurate, interpretable, and closely aligned with expert assessments, making it useful for improving automated radiology reporting.
2. **CheXbert [12]:** CheXbert evaluates the clinical accuracy of generated reports by treating evaluation as a multi-label classification task. It uses a BERT-based labeler to extract the presence, absence, or uncertainty of 14 clinical observations (e.g., Pneumonia, Cardiomegaly, No Finding) from both the generated and reference reports. We report the weighted F1 score between the two label sets, which quantifies the model's ability to correctly identify clinical findings regardless of specific phrasing.
3. **RadGraph [6]:** RadGraph assesses the factual and structural completeness of reports by parsing them into clinical knowledge graphs containing entities (e.g., anatomical structures, observations, pathologies) and relations (e.g., "located at," "suggestive of"). By computing the F1 score based on the overlap of entities and relations between the generated and reference graphs, this metric rewards models that correctly capture clinical dependencies and anatomical relationships, rather than isolated keywords.
4. **BERTScore [15]:** BERTScore evaluates the semantic similarity between generated and reference reports using token embeddings from a pre-trained language model. Unlike traditional metrics such as BLEU that rely on exact word matching, BERTScore computes cosine similarity between token representations, enabling it to recognize synonyms and paraphrases. This provides a measure of how well the overall meaning of the report are preserved

Multimodal Retrieval: Given a set of image–report pairs, we evaluate image-to-report retrieval using Recall@8 and Recall@16, which quantifies the proportion of test samples for which the correct report is retrieved among the top-8 or top-16 results, respectively. Retrieval is performed by computing the cosine similarity between image and text embeddings. Higher recall values indicate that the learned embedding space more effectively aligns visual and textual representations, enabling relevant image–report pairs to be retrieved more reliably.

B.4. Baseline Justification

We selected these baselines to provide a comprehensive comparison across both general-domain and medical-specific models.

1. **CheXpert Findings Classification:** We isolate and evaluate the quality of pretrained image embeddings from each model using linear probes on the 14 CheXpert findings.
   (a) **Chameleon [1]:** A 7B-parameter general-domain multimodal generative model trained on 4.2T image and text tokens. Included as the most comparable method to our generative pretraining approach, since it unifies images and text as tokens and trains autoregressively.
   (b) **HealthGPT [7]**: A medical-specific early-fusion multimodal generative model that integrates clinical images and text for comprehension and generation.
   (c) **RadPhi-2 [3]:** Text-only model pretrained on 2.7T tokens of medical text, including radiology reports; serves as a token prediction baseline without visual context.
   (d) **Masked Autoencoder (MAE) [5]:** Widely used masked image modeling baseline capturing strong visual representations; benchmarks a vision-only generative pretraining approach robust to image masking.
   (e) **M3AE [2]:** Multimodal masked autoencoder trained on chest X-rays and radiology reports, using a multimodal generative masking objective.
   (f) **CheXmix (S1):** Stage 1 model trained jointly on chest X-rays and radiology reports to assess the advantage of unified generative pretraining over text-only modeling.
   (g) **CheXmix (S1 + S2):** Stage 2 model building upon Stage 1 by introducing multimodal masked token prediction, allowing analysis of how masking improves classification performance.
   (h) **CheXagent [3]:** State-of-the-art multimodal large language model (MLLM) for chest X-rays, included as an upper-bound domain-specific baseline. We use the SigLIP image encoder pretrained on over 8 million chest X-rays.
2. **Image Inpainting:** Reconstruct full images from masked image tokens.
   (a) **VQ-GAN [1]:** Uses the Chameleon VQ-GAN image tokenizer to encode and decode masked image tokens; serves as a baseline for image reconstruction without full inpainting.
   (b) **RadPhi-2 [3]:** Text-only model; serves as a random token prediction baseline illustrating performance without visual information.
   (c) **CheXmix (S1):** Stage 1 model trained autoregressively on chest X-rays and radiology reports; establishes a baseline for reconstructing masked regions without explicit masked pretraining.
   (d) **CheXmix (S1 + S2):** Stage 2 model trained in a masked autoregressive manner; evaluates the impact of masked pretraining on image inpainting performance.
3. **Radiology Report Generation:** Generate radiology reports from chest X-rays.
   (a) **Chameleon [1]:** Evaluates the transferability of general-domain pretraining to radiology report generation.
   (b) **RadPhi-2 [3]:** Text-only model serving as a token prediction baseline without visual context for radiology reports.
   (c) **CheXmix (S1):** Stage 1 model trained autoregressively; generates reports by first inputting a chest X-ray image.
   (d) **CheXmix (S1 + S2):** Stage 2 model trained in a masked autoregressive manner; evaluates the ability to generate reports from image tokens.
   (e) **CheXagent [3]:** State-of-the-art MLLM for radiology report generation from chest X-rays.

C. Extended Quantitative Results

C.1. Extended Main Paper Results

	Generative Pretraining Objective						Reference Maximum
Masking %	**Chameleon**	**HealthGPT**	**MAE**	**M3AE**	**CheXmix (S1)**	**CheXmix (S1 + S2)**	**CheXagent**
0%	0.361 $\pm$0.000	0.319 $\pm$0.000	0.333 $\pm$0.001	0.359 $\pm$0.000	0.333 $\pm$0.000	<u>0.377 $\pm$0.000</u>	**0.520 $\pm$0.003**
20%	0.344 $\pm$0.001	0.301 $\pm$0.000	0.333 $\pm$0.002	<u>0.350 $\pm$0.001</u>	0.339 $\pm$0.001	0.344 $\pm$0.001	**0.404 $\pm$0.001**
40%	0.325 $\pm$0.000	0.247 $\pm$0.001	0.330 $\pm$0.004	0.351 $\pm$0.000	0.331 $\pm$0.001	**0.368 $\pm$0.001**	0.322 $\pm$0.000
60%	0.296 $\pm$0.000	0.243 $\pm$0.003	0.332 $\pm$0.003	0.344 $\pm$0.001	0.322 $\pm$0.000	**0.358 $\pm$0.002**	0.283 $\pm$0.001
80%	0.256 $\pm$0.000	0.217 $\pm$0.001	0.310 $\pm$0.003	0.327 $\pm$0.000	0.281 $\pm$0.001	**0.337 $\pm$0.001**	0.224 $\pm$0.001

Table 4. **CheXpert Embedding Findings Classification.** CheXmix (S1 + S2) demonstrates superior AUPRC performance at higher masking percentages compared to other generative baselines. **Bold** indicates the best-performing model for each masking percentage, while <u>underline</u> marks the best-performing generative model. AUPRC (mean $\pm$ std) is reported across three random seeds.

Masking %	**VQ-GAN**			**RadPhi-2**			**CheXmix (S1)**			**CheXmix (S1 + S2)**		
	PSNR↑	**MS-SSIM↑**	**FID↓**	**PSNR↑**	**MS-SSIM↑**	**FID↓**	**PSNR↑**	**MS-SSIM↑**	**FID↓**	**PSNR↑**	**MS-SSIM↑**	**FID↓**
10	15.00±0.01	0.671±0.00	236.5±0.75	22.55±0.04	0.856±0.00	78.8±0.73	**27.14±0.03**	**0.927±0.00**	**21.7±0.21**	26.95±0.04	0.920±0.00	27.8±0.24
20	10.61±0.01	0.482±0.00	335.4±0.81	18.98±0.04	0.735±0.00	160.6±1.21	**24.13±0.03**	**0.867±0.00**	**59.4±0.44**	**24.13±0.03**	0.859±0.00	75.3±0.46
30	8.39±0.01	0.385±0.00	422.6±1.11	15.84±0.05	0.644±0.00	227.0±1.37	22.04±0.03	**0.812±0.00**	**100.2±0.58**	**22.47±0.03**	0.809±0.00	116.5±0.67
40	7.18±0.01	0.329±0.00	427.0±0.93	13.47±0.05	0.582±0.00	264.9±1.21	19.78±0.03	0.757±0.00	**137.0±0.77**	**21.25±0.03**	**0.768±0.00**	141.1±0.89
50	6.59±0.02	0.288±0.00	402.3±1.01	12.03±0.05	0.531±0.00	288.0±1.16	16.87±0.02	0.690±0.00	179.1±0.80	**20.22±0.03**	**0.732±0.00**	**145.8±0.78**
60	6.29±0.02	0.252±0.00	415.6±1.05	11.08±0.05	0.485±0.00	321.5±1.11	13.84±0.02	0.607±0.00	233.2±0.95	**19.25±0.03**	**0.699±0.00**	**136.3±0.77**
70	6.06±0.02	0.237±0.00	468.7±0.95	10.52±0.04	0.447±0.00	364.2±1.08	11.53±0.02	0.520±0.00	283.8±1.04	**18.24±0.03**	**0.668±0.00**	**117.3±0.65**
80	5.91±0.02	0.249±0.00	514.5±1.04	10.38±0.04	0.418±0.00	376.7±1.20	10.24±0.02	0.455±0.00	302.3±1.10	**16.96±0.03**	**0.632±0.00**	**105.4±0.60**
90	5.74±0.02	0.275±0.00	468.2±1.04	10.64±0.03	0.388±0.00	391.6±1.15	9.57±0.02	0.407±0.00	284.1±1.04	**14.57±0.03**	**0.577±0.00**	**137.7±0.70**

Table 5. **Image Inpainting:** CheXmix (S1 + S2) improves inpainting performance, with the masked autoregressive model showing notable advantages at higher masking percentages (Best metrics are in **bold**). We compute PSNR, MS-SSIM, and FID on a random sample of 5,000 images and report mean and standard deviation across three runs with different random seeds.

GREEN Score

Mask %	Chameleon	RadPhi-2	CheXagent	CheXmix (S1)	CheXmix (S1 + S2)
0	0.019 ±0.005	0.014 ±0.004	0.152 ±0.011	0.219 ±0.015	**0.221 ±0.015**
10	0.017 ±0.005	0.005 ±0.002	0.146 ±0.011	0.200 ±0.015	**0.220 ±0.015**
20	0.039 ±0.007	0.002 ±0.001	0.134 ±0.009	0.181 ±0.013	**0.215 ±0.015**
30	0.012 ±0.003	0.003 ±0.002	0.133 ±0.010	0.147 ±0.012	**0.217 ±0.015**
40	0.022 ±0.004	0.002 ±0.001	0.117 ±0.010	0.148 ±0.012	**0.224 ±0.015**
50	0.015 ±0.004	0.005 ±0.002	0.112 ±0.009	0.122 ±0.012	**0.215 ±0.015**
60	0.017 ±0.004	0.003 ±0.002	0.103 ±0.009	0.073 ±0.009	**0.217 ±0.015**
70	0.019 ±0.006	0.003 ±0.002	0.085 ±0.008	0.052 ±0.008	**0.192 ±0.014**
80	0.015 ±0.005	0.006 ±0.002	0.068 ±0.008	0.051 ±0.006	**0.176 ±0.014**
90	0.010 ±0.003	0.010 ±0.003	0.052 ±0.007	0.037 ±0.008	**0.094 ±0.009**

CheXbert-F1

Mask %	Chameleon	RadPhi-2	CheXagent	CheXmix (S1)	CheXmix (S1 + S2)
0	0.202 ±0.018	0.154 ±0.018	0.383 ±0.022	**0.390 ±0.022**	**0.390 ±0.023**
10	0.164 ±0.018	0.156 ±0.017	**0.403 ±0.023**	0.396 ±0.022	0.398 ±0.023
20	0.178 ±0.018	0.154 ±0.019	0.357 ±0.023	**0.396 ±0.022**	**0.396 ±0.023**
30	0.170 ±0.019	0.131 ±0.018	0.354 ±0.023	0.389 ±0.022	**0.392 ±0.022**
40	0.171 ±0.019	0.140 ±0.019	0.343 ±0.022	0.376 ±0.022	**0.417 ±0.023**
50	0.182 ±0.019	0.115 ±0.017	0.325 ±0.021	0.380 ±0.022	**0.423 ±0.023**
60	0.202 ±0.020	0.129 ±0.017	0.278 ±0.021	0.359 ±0.022	**0.415 ±0.023**
70	0.182 ±0.019	0.098 ±0.015	0.269 ±0.020	0.374 ±0.022	**0.422 ±0.023**
80	0.184 ±0.019	0.100 ±0.015	0.185 ±0.017	0.343 ±0.021	**0.422 ±0.023**
90	0.166 ±0.019	0.122 ±0.016	0.112 ±0.015	0.283 ±0.022	**0.400 ±0.023**

RadGraph-F1

Mask %	Chameleon	RadPhi-2	CheXagent	CheXmix (S1)	CheXmix (S1 + S2)
0	0.026 ±0.003	0.006 ±0.002	0.091 ±0.007	**0.110 ±0.008**	0.104 ±0.008
10	0.022 ±0.002	0.002 ±0.001	0.090 ±0.006	0.101 ±0.008	**0.106 ±0.008**
20	0.032 ±0.003	0.001 ±0.001	0.083 ±0.006	0.087 ±0.007	**0.102 ±0.008**
30	0.017 ±0.002	0.000 ±0.000	0.073 ±0.006	0.073 ±0.006	**0.102 ±0.008**
40	0.013 ±0.002	0.001 ±0.001	0.068 ±0.006	0.068 ±0.005	**0.109 ±0.009**
50	0.016 ±0.002	0.000 ±0.000	0.062 ±0.005	0.045 ±0.004	**0.102 ±0.008**
60	0.014 ±0.002	0.000 ±0.000	0.048 ±0.005	0.032 ±0.004	**0.107 ±0.009**
70	0.014 ±0.002	0.000 ±0.000	0.045 ±0.005	0.024 ±0.003	**0.098 ±0.008**
80	0.014 ±0.002	0.000 ±0.000	0.033 ±0.004	0.027 ±0.003	**0.077 ±0.007**
90	0.009 ±0.002	0.000 ±0.000	0.030 ±0.004	0.008 ±0.003	**0.042 ±0.004**

BERTScore

Mask %	Chameleon	RadPhi-2	CheXagent	CheXmix (S1)	CheXmix (S1 + S2)
0	0.243 ±0.007	0.071 ±0.008	0.320 ±0.006	0.489 ±0.008	**0.503 ±0.008**
10	0.241 ±0.007	0.035 ±0.009	0.318 ±0.006	0.483 ±0.009	**0.505 ±0.008**
20	0.253 ±0.007	0.016 ±0.006	0.310 ±0.005	0.477 ±0.008	**0.499 ±0.008**
30	0.213 ±0.009	0.007 ±0.006	0.306 ±0.005	0.455 ±0.009	**0.506 ±0.007**
40	0.161 ±0.010	0.002 ±0.011	0.305 ±0.005	0.463 ±0.007	**0.508 ±0.008**
50	0.175 ±0.009	0.079 ±0.017	0.295 ±0.005	0.430 ±0.007	**0.500 ±0.008**
60	0.155 ±0.009	0.044 ±0.005	0.291 ±0.005	0.394 ±0.009	**0.505 ±0.009**
70	0.154 ±0.009	0.071 ±0.005	0.287 ±0.005	0.381 ±0.009	**0.501 ±0.008**
80	0.164 ±0.009	0.054 ±0.006	0.271 ±0.005	0.360 ±0.008	**0.480 ±0.008**
90	0.131 ±0.013	0.029 ±0.006	0.263 ±0.005	0.272 ±0.009	**0.450 ±0.007**

Table 6. **Radiology report generation.** CheXmix (S1 + S2) achieves the best performance across masking percentages (best metrics in **bold**). We compute GREEN score, CheXbert-F1, RadGraph-F1, and BERTScore on a random sample of 1,000 radiology reports and report mean and standard deviation across three runs with different random seeds.

Mask %	CheXmix S1		CheXmix S1 + S2	
	B (No CM)	**CM** (Causal)	**B** (No CM)	**CM** (Causal)
a) CheXpert Classification				
AUROC (↑)				
0	0.713 ±0.000	0.664 ±0.000	**0.716 ±0.000**	0.712 ±0.000
20	0.702 ±0.000	0.660 ±0.000	0.684 ±0.000	**0.705 ±0.000**
40	0.678 ±0.000	0.653 ±0.001	0.691 ±0.000	**0.702 ±0.000**
60	0.667 ±0.000	0.643 ±0.000	0.682 ±0.000	**0.689 ±0.000**
80	0.591 ±0.001	0.624 ±0.000	0.632 ±0.000	**0.656 ±0.000**
AUPRC (↑)				
0	**0.382 ±0.000**	0.333 ±0.000	0.364 ±0.000	0.377 ±0.000
20	**0.364 ±0.001**	0.339 ±0.001	0.336 ±0.001	0.344 ±0.001
40	0.340 ±0.001	0.331 ±0.001	0.345 ±0.000	**0.368 ±0.001**
60	0.325 ±0.000	0.322 ±0.000	0.343 ±0.000	**0.358 ±0.002**
80	0.277 ±0.001	0.281 ±0.001	0.301 ±0.000	**0.337 ±0.001**
b) Image Inpainting				
PSNR (↑)				
20	23.30 ±0.01	24.13 ±0.03	**24.57 ±0.03**	24.13 ±0.03
40	18.94 ±0.01	19.78 ±0.03	**21.93 ±0.03**	21.25 ±0.03
60	14.65 ±0.01	13.84 ±0.02	**20.30 ±0.03**	19.25 ±0.03
80	11.57 ±0.01	10.24 ±0.02	**18.54 ±0.03**	16.96 ±0.03
MS-SSIM (↑)				
20	0.835 ±0.00	**0.867 ±0.00**	0.863 ±0.00	0.859 ±0.00
40	0.675 ±0.00	0.757 ±0.00	**0.775 ±0.00**	0.768 ±0.00
60	0.500 ±0.00	0.607 ±0.00	**0.713 ±0.00**	0.699 ±0.00
80	0.346 ±0.00	0.455 ±0.00	**0.656 ±0.00**	0.632 ±0.00
FID (↓)				
20	107.1 ±0.8	**59.4 ±0.4**	77.7 ±0.5	75.3 ±0.5
40	251.3 ±0.9	**137.0 ±0.8**	163.9 ±0.8	141.1 ±0.9
60	314.5 ±1.0	233.2 ±1.0	154.6 ±0.8	**136.3 ±0.8**
80	332.3 ±1.1	302.3 ±1.1	**69.50 ±0.6**	105.4 ±0.6
c) Report Generation				
GREEN Score (↑)				
0	0.172 ±0.02	0.219 ±0.02	0.140 ±0.01	**0.221 ±0.02**
20	0.153 ±0.01	0.181 ±0.01	0.141 ±0.01	**0.215 ±0.01**
40	0.112 ±0.01	0.148 ±0.01	0.132 ±0.01	**0.224 ±0.01**
60	0.071 ±0.01	0.073 ±0.01	0.129 ±0.01	**0.217 ±0.01**
80	0.061 ±0.01	0.051 ±0.01	0.104 ±0.01	**0.176 ±0.01**
CheXbert F1 (↑)				
0	0.358 ±0.02	**0.390 ±0.02**	0.300 ±0.02	**0.390 ±0.02**
20	0.336 ±0.02	**0.396 ±0.02**	0.305 ±0.02	**0.396 ±0.02**
40	0.343 ±0.02	0.376 ±0.02	0.300 ±0.02	**0.417 ±0.02**
60	0.339 ±0.02	0.359 ±0.02	0.304 ±0.02	**0.415 ±0.02**
80	0.335 ±0.02	0.343 ±0.02	0.305 ±0.02	**0.422 ±0.02**
RadGraph F1 (↑)				
0	0.086 ±0.01	**0.110 ±0.01**	0.063 ±0.01	0.104 ±0.01
20	0.071 ±0.01	0.087 ±0.01	0.062 ±0.01	**0.102 ±0.01**
40	0.054 ±0.01	0.068 ±0.01	0.062 ±0.01	**0.109 ±0.01**
60	0.024 ±0.01	0.032 ±0.00	0.059 ±0.01	**0.107 ±0.01**
80	0.012 ±0.00	0.027 ±0.00	0.040 ±0.01	**0.077 ±0.01**
BERTScore (↑)				
0	0.483 ±0.01	0.489 ±0.01	0.437 ±0.01	**0.503 ±0.01**
20	0.468 ±0.01	0.477 ±0.01	0.439 ±0.01	**0.499 ±0.01**
40	0.448 ±0.01	0.463 ±0.01	0.432 ±0.01	**0.508 ±0.01**
60	0.458 ±0.01	0.394 ±0.01	0.430 ±0.01	**0.505 ±0.01**
80	0.454 ±0.01	0.360 ±0.01	0.400 ±0.01	**0.480 ±0.01**

Table 7. **Extended Causal Mask Ablation.** We evaluate CheXmix pretrained with 50% masking using either bidirectional attention (B) or a causal mask (CM) across three tasks: (a) classification, (b) image inpainting, and (c) report generation. For both classification and report generation, CheXmix S1+S2 with CM consistently achieves the strongest performance across masking ratios. Although CheXmix S1+S2 (B) performs slightly better on image inpainting metrics, the differences between B and CM are small, typically only a few percentage points in PSNR and MS-SSIM.

Model	GREEN↑	CheXbert↑	RadGraph-F1↑	BERTScore↑
Chameleon	0.0287 ±0.001	0.139 ±0.017	0.033 ±0.003	0.272 ±0.005
HealthGPT	0.216 ±0.017	0.239 ±0.021	0.075 ±0.004	0.398 ±0.005
CheXmix (S1 + S2)	**0.217 ±0.016**	**0.413 ±0.027**	**0.076 ±0.006**	**0.426 ±0.007**

Table 8. **ReXGradient-160K Report Generation (External Validation).** CheXmix (S1 + S2) outperforms early-fusion generative models on the ReXGradient-160K report generation task. The metrics are GREEN, CheXbert, RadGraph-F1, and BERTScore on a random sample of 500 radiology reports from the test set. There is no masking used for this experiment. We report the mean and standard deviation across three runs with different random seeds. CheXagent had metrics of GREEN: 0.135 ±0.011, CheXbert: 0.310 ±0.024, RadGraph: 0.068 ±[std], BERTScore: 0.398 ±0.005 on this task.

C.2. Extended Ablations

Masking Ratio Ablation: We conduct linear probe experiments on the CheXpert findings classification task using embeddings pretrained with four masking ratios (25%, 50%, 75%, and 90%) over 100K steps to evaluate how masking affects representation quality (Table 9). We find that 50% masking yields the highest AUROC and AUPRC, and therefore select this ratio for CheXmix (S1 + S2) pretraining. Higher masking ratios (75% and 90%) produce comparable classification performance, with differences within 0.01 AUROC. While general-domain studies have reported benefits from higher masking ratios [5], prior work on chest X-ray autoencoders suggests that lower masking ratios can be more effective [14].

Mask Pct (%)	AUROC	AUPRC
25	0.672	0.337
50	**0.676**	**0.341**
75	0.669	0.334
90	0.641	0.327

Table 9. **Masking Ratio Ablation.** Effect of different masking percentages during CheXmix (S1 + S2) pretraining on CheXpert classification performance. AUROC and AUPRC are reported for each masking ratio, showing that 50% masking yields the best discriminative performance. Consequently, we pretrain CheXmix (S1 + S2) with 50% masking.

CheXmix S1 Extended: CheXmix (S1) refers to an intermediate checkpoint obtained after Stage 1 training, without proceeding to Stage 2. On CheXpert classification, CheXmix S1-extended trained for additional S1 steps achieves similar AUROC performance compared to the original S1 checkpoint (0.667±0.001 vs. 0.664±0.000). Notably, despite having the same total number of training steps, CheXmix (S1 + S2) outperforms CheXmix S1-extended by 6.75%, highlighting that Stage 2 multimodal masked training improves representation quality beyond simply increasing training duration.

CheXmix Variant	Total Training Steps	AUROC
CheXmix S1	703,671	0.664 ± 0.000
CheXmix S1-extended	1,217,664	0.667 ± 0.001
CheXmix (S1 + S2)	1,217,664	$\mathbf{0.712 \pm 0.001}$

Table 10. **S1 Extended Ablation (no masking).** Comparison of CheXmix variants showing that additional Stage 1 training (S1-extended) yields minimal gains, while Stage 2 multimodal masked training strategy (S1 + S2) improves AUROC at the same total training budget.

D. Extended Qualitative Results

D.1. Image Inpainting Examples

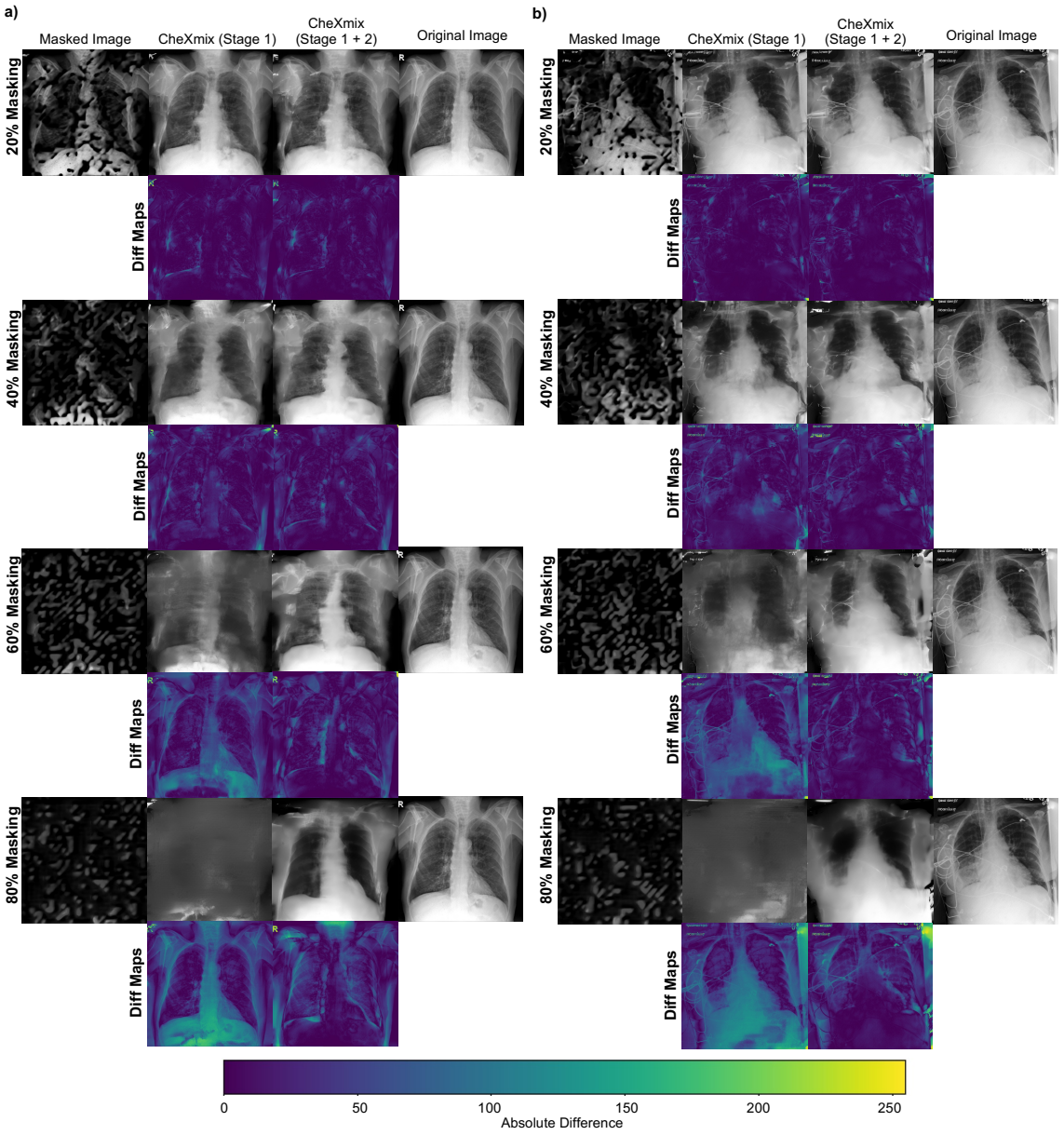

Figure 1. **Image Inpainting Visualization** CheXmix (S1 + S2) pretraining provides substantial improvements in inpainting quality at higher masking ratios for (a) a chest radiograph with COPD and (b) a case with consolidation.

D.2. Radiology Report Examples

Model	Generated Report
	Sample 1: Normal Case
Reference	*Findings:* The lungs are well inflated and clear. Heart size and mediastinal contours are normal. There is no pleural effusion or pneumothorax. Osseous structures are intact. *Impression:* No acute cardiopulmonary process.
Chameleon	This is a posteroanterior chest radiograph, which is a type of X-ray image that is taken from the back and shows the front of the chest. The image appears to be of a patient with a medical condition, as evidenced by the presence of a tube or catheter in the right upper lobe of the lung, which is likely being used to treat a respiratory issue. The tube is attached to a device that is emitting a white, glowing light, which suggests that the tube is being used for some sort of medical procedure. The patient's heart is also visible on the X-ray, and it appears to be in good condition. The lungs are not fully inflated, which may indicate that the patient is experiencing some difficulty breathing or that they have recently been intubated. The ribs are also visible on the image, and they appear to be in good condition. The spine is also visible, and it appears to be straight and well-aligned. Overall, the image appears to show a patient with a medical condition that is being treated with a tube or catheter in the right upper lobe of the lung.
CheXagent	[Cardiac: Heart Size] No acute cardiopulmonary process.
CheXmix (S1)	*Findings:* The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are stable. *Impression:* No acute cardiopulmonary process. No significant interval change. No new focal
CheXmix (S1 + S2)	*Findings:* The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable. *Impression:* No acute cardiopulmonary process. No focal consolidation to suggest pneumonia.
	Sample 2: Case with pleural effusions and support devices
Reference	*Findings:* In comparison with the study of ___, the left chest tube and subclavian catheters remain in place. No evidence of pneumothorax. Cardiac silhouette is at the upper limits of normal in size. There is evidence of elevated pulmonary venous pressure. Mild haziness of the lower zones is consistent with small bilateral pleural effusions and compressive atelectasis at the bases. *Impression:*
Chameleon	The image depicts a chest x-ray of a patient with a suspected pulmonary embolism (PE). The findings on the x-ray include: • Right lower lobe consolidation with air bronchograms, suggesting an acute pulmonary embolism (PE) • A large pleural effusion present on the right side, which could be related to the PE or a separate condition • A small pleural effusion present on the left side, which may be related to the PE or a separate condition • Cardiomegaly, which is suggestive of heart failure or other cardiac conditions The impression based on the findings is that the patient likely has a pulmonary embolism (PE), as well as a possible cardiomyopathy or heart failure. The large pleural effusions on both sides may be related to the PE or a separate condition. Further testing, such as a CT pulmonary angiogram (CTPA) or ventilation-perfusion (VQ) scan, may be necessary to confirm the diagnosis and guide treatment.
CheXagent	[Breathing: Lungs] Mild pulmonary edema is present. [Breathing: Pleura] A small right pleural effusion is noted. [Breathing: Lungs] Mild bibasilar atelectasis is observed.
CheXmix (S1)	*Findings: Impression:* AP chest compared to ___: Previous mild pulmonary edema has improved, but there is still substantial bibasilar atelectasis, left greater than right, and small pleural effusionss, left greater than right. Heart size is normal. ET tube and right internal jugular line are in standard positions.
CheXmix (S1 + S2)	*Findings: Impression:* In comparison with the study of ___, there is little change in the appearance of the heart and lungs. Again there is enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases. The right IJ catheter extends to the mid portion of the SVC. No evidence of pneumothorax.

Table 11. **Qualitative comparison of generated reports.** Sample 1 demonstrates a normal case. Sample 2 demonstrates a complex case with support devices and pleural effusions. Generated reports from CheXmix (S1+S2) are shown for comparison.

E. Test-Time Augmentation Prompt

```
You are an expert medical report synthesizer. Your task is to analyze
five generated radiology reports below. These reports are variations
describing the same patient.

Compile them into one, single, consolidated report that contains:
1. A single ``Findings'' section.
2. A single ``Impressions'' section.

Rules:
1. Combine and de-duplicate all repetitive information.
2. Make sure the synthesized report is the SAME LENGTH as the original
   reports.
3. If there are slight variations in wording for the same finding, use
   the most precise and complete description.
4. Ensure the final ``Findings'' and ``Impressions'' are comprehensive
   and written as a single, coherent section each with no newlines or
   bullet points.
5. If no findings or impressions are present in the generated reports,
   then leave blank.
```

References

[1] Team Chameleon. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3, 4, 6, 9

[2] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 9

[3] 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, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. 9

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 6, 9, 13

[6] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021. 4, 8

[7] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation, 2025. 9

[8] Rishi P Mathew, Timothy Alexander, Vimal Patel, and Gavin Low. Chest radiographs of cardiac devices (part 1): Lines, tubes, non-cardiac medical devices and materials. *SA Journal of Radiology*, 23(1):1–9, 2019. 3

[9] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology

report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390, 2024. 4, 8

[10] Ehsan Samei, Michael J Flynn, Edward Peterson, and William R Eyler. Subtle lung nodules: influence of local anatomic variations on detection. *Radiology*, 228(1):76–84, 2003. 3

[11] Chandrakant Manmath Shetty, Ashita Barthur, Avinash Kambadakone, Nilna Narayanan, and Rajagopal Kv. Computed radiography image artifacts revisited. *American Journal of Roentgenology*, 196(1):W37–W47, 2011. 3

[12] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020. 4, 8

[13] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 6

[14] Xin Xing, Gongbo Liang, Chris Wang, Nathan Jacobs, and Ai-Ling Lin. Self-supervised learning application on covid-19 chest x-ray image classification using masked autoencoder. *Bioengineering*, 10(8):901, 2023. 6, 13

[15] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 4, 8

[16] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024. 3

[17] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multimodal model. *arXiv preprint arXiv:2408.11039*, 2024. 6