CHEXPERCEPT: A BENCHMARK FOR EVALUATING EXPERT-LEVEL LESION PERCEPTION IN CHEST X-RAYS

Anonymous authors
Paper under double-blind review

ABSTRACT

The evaluation of vision-language models (VLMs) for chest X-ray (CXR) analysis has largely been limited to disease-presence classification without visual grounding. Such evaluations fail to verify the expert-level lesion perception necessary to ensure the clinical reliability of VLMs. To address these limitations, we introduce **CheXpercept**, a sequential, multi-level perception benchmark that mirrors a radiologist's cognitive workflow across coarse-level detection, fine-level contour evaluation and revision, and semantic-level attribute extraction. To ensure high clinical fidelity at scale, we construct the dataset using a semi-automated generation pipeline paired with review by six medical experts. CheXpercept contains 12,000 QA items derived from 2,400 CXRs, covering eight clinically critical pulmonary and cardiac lesions. To demonstrate the current landscape of VLM perception, we benchmark 18 general and medical VLMs on CheXpercept. The models achieve adequate performance only at the coarse level, with accuracy degrading precipitously on deeper visual tasks. Notably, recent general models outperform recent medical VLMs overall. While medical models show marginal improvements over their specific backbones at certain perception levels, the inherent capability of the backbone itself ultimately dictates performance far more than current domain adaptation strategies.

1 INTRODUCTION

Radiology is central to modern medicine, providing the visual evidence that drives clinical diagnosis and treatment planning. Within this field, the chest X-ray (CXR) is the most accessible modality for evaluating cardiopulmonary conditions (Raoof et al., 2012). A single CXR encodes a broad range of clinical information, among which pulmonary and cardiac lesions represent the primary clinical findings that radiologists describe at length in their reports. In particular, pulmonary lesions (*e.g.*, pneumonia) present significant challenges because they appear at varying locations with irregular shapes, while cardiac lesions (*e.g.*, cardiomegaly) are difficult to demarcate precisely due to overlapping opacities and anatomical structures. To analyze such complex lesions, radiologists interpret a CXR through a sequential cognitive workflow (Bhattacharya et al., 2022; Wu et al., 2026) across three perception levels: *coarse-level*, identifying the presence of abnormalities; *fine-level*, delineating lesion contours; and *semantic-level*, extracting lesion attributes (*e.g.*, severity) that form the basis of the report. Crucially, a perceptual error at any stage cascades, affecting not only subsequent perceptions but also final decisions. This demands careful attention at each stage, rendering the overall workflow inherently labor-intensive (Degnan et al., 2019).

Recent advances in vision-language models (VLMs) have spurred efforts to automate radiology workflows through visual question answering (VQA) (Li et al., 2023; Sellergren et al., 2025; 2026) and report generation (Bannur et al., 2024; Hyland et al., 2023; Myronenko et al., 2025). Current VLMs seem capable of answering complex questions and generating detailed descriptions regarding lesions. However, it remains unclear whether these capabilities actually stem from expert-level visual perception (*i.e.*, operating across all three perception levels), or from surface-level image-text association. To establish clinical reliability, we must therefore verify that a VLM perceives lesions at a radiologist's level of granularity. Despite this necessity, current CXR benchmarks struggle to evaluate fine-grained perception for pulmonary and cardiac lesions. Most evaluations ask only

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

whether a disease is present, effectively probing *coarse-level* perception without any visual grounding (Ben Abacha et al., 2021; Lau et al., 2018; Zhang et al., 2023). Even benchmarks with visual grounding fall short: some rely on coarse bounding boxes, while others leverage organ segmentation masks yet limit evaluation to simple tasks such as anatomy detection or linear measurements (Lee et al., 2025). At present, no benchmark assesses how accurately a VLM perceives clinically critical pulmonary and cardiac lesions across the full perceptual workflow.

To address this gap, we introduce **CheXpercept**, a multi-level perception benchmark designed to evaluate VLMs across a multi-stage interpretation workflow anchored to lesion segmentation masks. CheXpercept evaluates a VLM along three perception levels: (i) *coarse-level* perception, detecting the presence of a lesion; (ii) *fine-level* perception, evaluating and revising a candidate lesion contour; and (iii) *semantic-level* perception, extracting four lesion attributes (distribution, location, severity, and comparison) (Moon et al., 2025). To ensure high clinical fidelity without sacrificing scale, CheXpercept is constructed via a semi-automated pipeline. This approach pairs automated QA generation with continuous review and final verification by six medical experts, alleviating the manual annotation bottleneck while retaining full physician oversight. The resulting benchmark comprises eight major lesion types, 2,400 CXRs, and 12,000 QA items.

Through benchmarking of 18 VLMs, CheXpercept reveals critical perceptual limitations in current models. While most models perform adequately at coarse-level perception, accuracy collapses once fine-level or semantic-level perception is required. Strikingly, while recent medical VLMs show only marginal improvements over their foundation backbones at specific levels, they generally fail to outperform current general VLMs at fine and semantic levels. This highlights a critical limitation: current medical domain adaptation strategies fail to meaningfully enhance fine-grained visual perception.

2 Related works

Early CXR benchmarks (Ben Abacha et al., 2021; Lau et al., 2018; Zhang et al., 2023) primarily focused on coarse-level perception, evaluating only the presence and type of abnormalities. In contrast, recent benchmarks (Bae et al., 2024; Chen et al., 2024b; Fallahpour et al., 2025; Hu et al., 2023; Liu et al., 2025; Pal et al., 2025; Zuo et al., 2025) have shifted toward high-level diagnostic and linguistic reasoning evaluation. However, such reasoning benchmarks evaluate only the final answer in a single stage, conflating perception and reasoning into an *entangled* metric. Furthermore, visual grounding is largely absent across these benchmarks. Although a few benchmarks (Chen et al., 2024b; Liu et al., 2025) incorporate bounding box annotations, these localizations remain too coarse to evaluate a model's ability to delineate precise lesion contours.

More recently, CXReasonBench (Lee et al., 2025) introduced a multi-stage format that attempts to verify anatomy-level perception by utilizing organ segmentation masks (*e.g.*, aorta, mediastinum). However, relying solely on such masks inherently restricts its scope to anatomically defined conditions (*e.g.*, aortic enlargement, mediastinal widening). Moreover, its evaluation is confined to anatomy recognition and geometric measurements (*e.g.*, width estimation), leaving the fine-grained perception of major pulmonary and cardiac lesions unaddressed. To overcome these structural limitations, CheXpercept is designed to decouple perception from reasoning more explicitly. By providing a sequential evaluation framework with pulmonary and cardiac lesion segmentation masks, our benchmark enables more precise verification of perception across clinically important lesions in CXRs. A comparison is provided in Table 1.

3 CheXpercept

CheXpercept mimics the multi-level perception workflow of radiologists, dissecting a VLM's visual capabilities across three distinct levels and four sequential stages (§3.1). To reflect realistic clinical scenarios, the evaluation dynamically branches into three paths based on lesion presence and segmentation mask quality (§3.2). The benchmark targets eight major lesions that are most frequently mentioned in radiology reports: cardiomegaly, pneumonia, pneumothorax, atelectasis, opacity, consolidation, edema, and effusion. By allocating 100 QA sequences per (lesion, path) combination, CheXpercept comprises 2,400 CXRs and 12,000 QA items in total. An overview of the full sequence is shown in Figure 1.

Table 1: Comparison of medical VQA benchmarks. Since existing benchmarks are not explicitly designed for lesion perception, we analyze the extent to which their constituent questions correspond to three perception levels: coarse (lesion presence), fine (lesion contour), and semantic (lesion attributes). △ indicates that only a subset of questions partially overlaps with semantic-level tasks (*e.g.*, benchmarks that include only measurement tasks for lesion size). QA formats abbreviated as SS = Single-Stage, MS = Multi-Stage, MP = Multi-Path.

Benchmark	VQA Pairs	QA Format	Lesion Perception			Visual Grounding
			Coarse	Fine	Semantic	
VQA-Med (Ben Abacha et al., 2021)	5.5K	SS	✓	✗	✗	✗
VQA-RAD Lau et al. (2018)	3.5K	SS	✓	✗	△	✗
PMC-VQA (Zhang et al., 2023)	227K	SS	✓	✗	△	✗
MIMIC-CXR-VQA (Bae et al., 2024)	377K	SS	✓	✗	△	✗
MIMIC-Diff-VQA (Hu et al., 2023)	701K	SS	✓	✗	△	✗
ChestAgentBench (Fallahpour et al., 2025)	2.5K	SS	✓	✗	△	✗
MedXpertQA (Zuo et al., 2025)	2K	SS	✓	✗	△	✗
ReXVQA (Pal et al., 2025)	696K	SS	✓	✗	△	✗
GEMeX (Liu et al., 2025)	1.6M	SS	✓	✗	△	✓ (bbox)
CheXbench (Chen et al., 2024b)	8K	SS	✓	✗	△	✓ (bbox)
CXReasonBench (Lee et al., 2025)	19K	MS, MP	✓	✗	△	✓ (organ mask)
CheXpercept (Ours)	12K	MS, MP	✓	✓	✓	✓ (lesion mask)

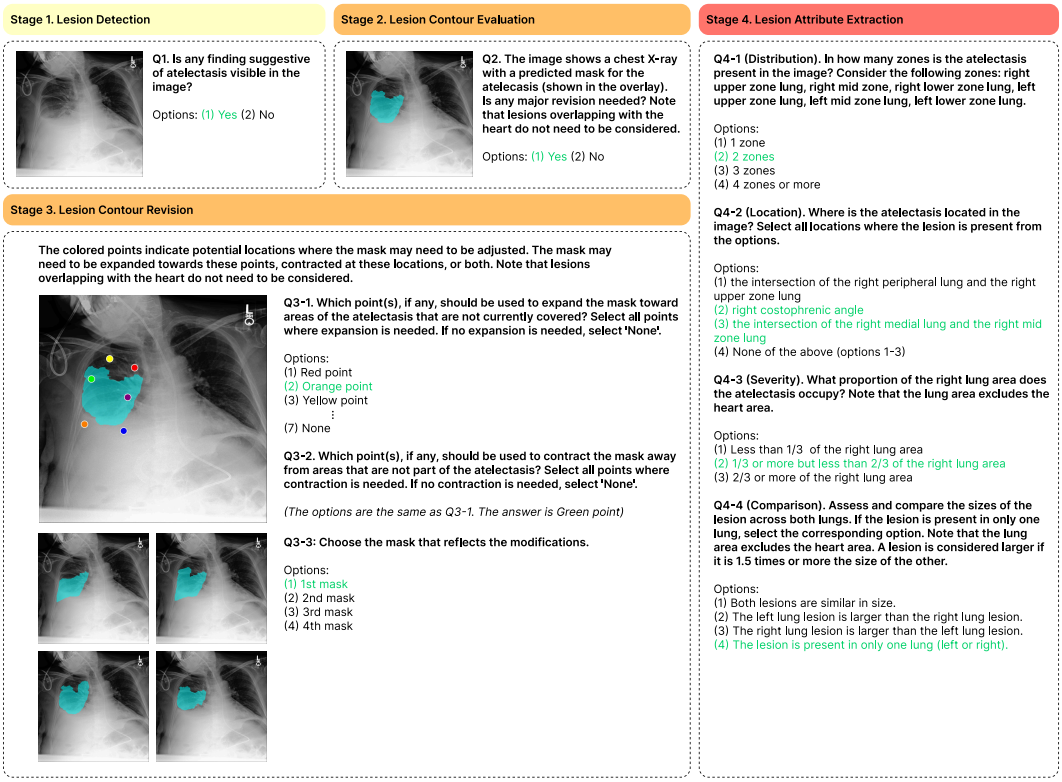


Figure 1: Overview of the CheXpercept benchmark. The evaluation systematically advances from basic lesion detection (Stage 1) to spatial contour evaluation and revision (Stages 2 and 3), culminating in detailed clinical attribute extraction (Stage 4).

3.1 Stages and Perception Levels

3.1.1 Stage 1: Lesion Detection (Coarse-Level)

The first stage assesses the model's screening ability to determine whether a target lesion is present in a CXR. Given a raw CXR and a question specifying a lesion type, the model outputs a binary ("Yes" or "No") response regarding the presence of the lesion. The question phrasing varies by lesion type: since opacity and consolidation are direct radiographic findings, they are queried with "Is there

any `[lesion]` visible in the image?"; conversely, the remaining lesions require clinical inference combining visual signs with other patient data (*e.g.*, symptoms). Since the model relies solely on the image, these are queried with "Is any finding suggestive of `[lesion]` visible in the image?"

3.1.2 STAGE 2: LESION CONTOUR EVALUATION (FINE-LEVEL)

Moving beyond binary detection, this stage evaluates the model's ability to judge the precise boundaries of a target lesion. With a candidate lesion mask overlaid on the CXR, the model is required to answer a binary question about whether the mask requires major revision. The candidate mask is either an *optimal* lesion mask, derived directly from the ground truth segmentation, for which the correct answer is "No", or a *suboptimal* lesion mask, generated by perturbing the optimal mask, for which the correct answer is "Yes". The construction of suboptimal masks is detailed in §4.4.

3.1.3 STAGE 3: LESION CONTOUR REVISION (FINE-LEVEL)

Following the evaluation stage, this task assesses whether the model can actively refine a suboptimal mask toward the true lesion boundary, mirroring the clinical training process in which medical trainees iteratively revise their segmentation attempts under faculty supervision. To circumvent the structural limitations of current VLMs in direct mask generation, and to avoid the excessive context length introduced by iterative revision scenarios, we decompose this task into two types of multiple-choice questions that together require the model to complete all revisions in a single attempt:

- **Point-wise revision**: Two consecutive queries are issued over the same visual prompt, in which up to eight color-coded points are placed near the suboptimal mask boundary. The model is first asked to select the points at which the mask should be expanded, and then the points at which it should be contracted.
- **Revision result selection**: Four candidate masks are presented, and the model selects the one that best reflects the previously identified points for expansion and contraction.

3.1.4 STAGE 4: LESION ATTRIBUTE EXTRACTION (SEMANTIC-LEVEL)

The final stage evaluates whether the model can extract essential lesion attributes required for writing radiology reports. With the optimal mask overlaid, the model answers multiple-choice questions regarding four semantic attributes (Moon et al., 2025). To eliminate ambiguous qualitative descriptions, each attribute is grounded in explicit quantitative criteria:

- **Distribution** (*e.g.*, diffuse, multifocal): The spatial spread of a lesion across the lungs, operationalized as the exact number of lung zones occupied by the lesion.
- **Location**: The anatomical position of the lesion, identified through a multi-select question over 20 predefined regions that overlap with the mask.
- **Severity** (*e.g.*, mild, severe): The clinical burden of a lesion, computed as the ratio of the lesion area to the total lung area. This approach follows standard radiological practice where lesion size serves as the primary determinant of severity, with the ratio discretized into three intervals using thresholds of $1/3$ and $2/3$.
- **Comparison** (*e.g.*, predominant on the right): The relative predominance across the lungs; a side is considered predominant if its lesion area is at least 1.5 times that of the contralateral side.

3.2 EVALUATION PATHS

To reflect realistic clinical scenarios where lesion presence and mask quality vary, the evaluation dynamically branches into three distinct paths based on lesion presence in Stage 1 and mask quality in Stage 2. Importantly, as the stages progress, all context from previous stages is retained:

- **Revision-Required (RR)**: The lesion is present and the candidate mask is suboptimal, so the model must traverse the full pipeline of detection, contour evaluation, revision, and attribute extraction (Stage $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$).
- **Revision-Free (RF)**: The lesion is present and the candidate mask is already optimal, so the model is expected to recognize the mask as correct and skip the revision stage (Stage $1 \rightarrow 2 \rightarrow 4$).

- **Lesion-Free (LF)**: The target lesion is absent, so the workflow terminates immediately after the initial screening (Stage 1 only).

Cardiomegaly serves as a structural exception; as lung-based semantic attributes in Stage 4 are inapplicable, the RR path simplifies to Stage $1 \rightarrow 2 \rightarrow 3$, and the RF path to Stage $1 \rightarrow 2$. The distribution of gold answers across all paths and stages is detailed in Appendix A.3.

4 SEMI-AUTOMATED BENCHMARK GENERATION

A recurring challenge in constructing medical benchmarks is the reliance on manual expert annotation, which often becomes a significant bottleneck that hinders scalability. To address this, CheXpercept employs a semi-automated framework designed to minimize manual labor while maintaining expert-level quality. The pipeline consists of five stages: (i) construction of candidate pools for lesion masks and normal CXRs; (ii) expert selection of *optimal* masks and *true-normal* CXRs; (iii) extraction of geometric information from the selected masks; (iv) automated mask deformation to generate suboptimal counterparts; and (v) automated QA generation followed by a final expert validation.

4.1 CANDIDATE POOL CONSTRUCTION

CheXpercept requires a large collection of CXRs and lesion masks. To this end, we leverage MIMIC-ILS (Choi et al., 2025; 2026) and SIIM-ACR Pneumothorax Segmentation (Zawacki et al., 2019). The datasets provide CXRs alongside binary labels for lesion presence and the corresponding segmentation masks. Using these resources, we establish two separate data pools: one collection of abnormal CXRs with their corresponding lesion masks, and another pool of normal CXRs where target lesions are absent (details in Appendix B.2).

4.2 EXPERT CURATION OF OPTIMAL MASKS AND TRUE-NORMAL CXRS

Since MIMIC-ILS was constructed through an automated framework that targets clinically acceptable quality, its mask precision is insufficient for the contour-level evaluation required by Stages 2 and 3 in CheXpercept. Furthermore, a small fraction of CXRs labeled as normal may contain subtle lesions identifiable only upon rigorous inspection. To guarantee ground-truth integrity, a panel of six medical experts conducted a manual review of the candidate pool. They retained only those masks that precisely trace the lesion boundaries as *optimal* masks, and confirmed lesion-free images as *true-normal* CXRs. In contrast, both the lesion masks and normal CXRs from the SIIM-ACR dataset were originally annotated and verified by radiologists; therefore, they were directly adopted as *optimal* masks and *true-normal* CXRs without additional review. These combined resources are used to construct RR and RF path items, while true-normal CXRs form the basis for LF path items. Because experts only have to review candidates rather than draw masks from scratch, the annotation cost is drastically reduced.

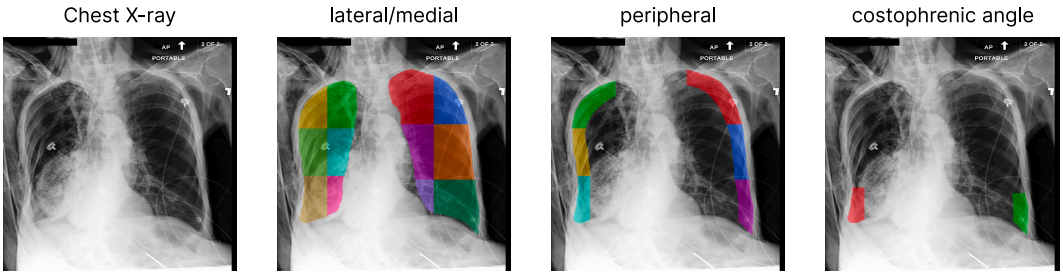


Figure 2: Visualization of the 20 lung sub-regions. They are formed by intersecting the basic upper, middle, and lower zones with lateral, medial, and peripheral regions. The costophrenic angle is additionally delineated as a separate zone.

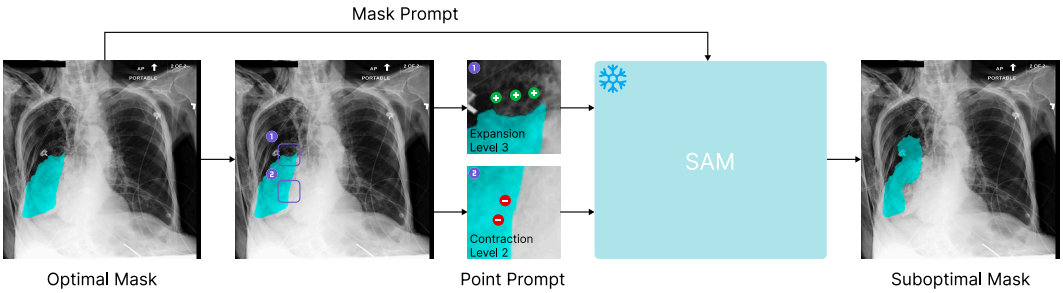


Figure 3: Overview of the suboptimal mask generation process. First, disjoint lung sub-regions overlapping with the optimal mask are selected, and expansion or contraction operations are assigned to each with varying magnitudes. Point prompts are then sampled within these designated regions to guide the deformation. Finally, the optimal mask and the sampled point prompts are provided as prompts to SAM3, yielding a refined suboptimal mask.

4.3 Geometric Information Extraction

We extract diverse geometric properties, including anatomical location and size, from the curated optimal masks. To characterize lesion location at the granularity used in real radiology reports and to enable downstream suboptimal mask generation, we map each CXR to a rich spatial representation. Building upon lung masks produced by a pretrained HybridGNet Cosarinsky et al. (2025), we employ a custom partitioning algorithm, co-designed with medical experts, to divide the lungs into 20 fine-grained sub-regions that reflect standard anatomical references (Figure 2; details in Appendix B.4). By leveraging these global and multi-region masks, we automatically derive the ground truth labels for the semantic attribute questions in Stage 4.

4.4 Suboptimal Mask Generation

To construct RR-path items, we generate *suboptimal* lesion masks that contain intentional errors for use in Stages 2 and 3. We design an automated mask deformation framework built on SAM3 Carion et al. (2025). Although SAM3 was originally a promptable segmentation model designed to delineate objects using point and mask prompts, we repurpose it as a precise mask *deformer*. Specifically, the expert-curated *optimal* mask serves as the initial mask input, while automatically sampled point prompts at targeted locations actively warp the contour.

To produce suboptimal masks that resemble realistic clinical errors, we carefully control both the position and the number of point prompts (Figure 3): the direction of deformation (expansion or contraction) is determined by prompt polarity, the magnitude by prompt count, and these operations are confined to disjoint sub-regions to prevent geometric interference (full mechanics in Appendix B.5).

To ensure diversity within the benchmark, we apply a combination of up to three deformation operations to a single optimal mask, producing a family of derivatives. For each RR-path, one derivative is presented in Stage 2 to test the model's contour recognition. The remaining derivatives serve as plausible distractors in the Stage 3 revision result selection, rigorously evaluating the model's ability to discern subtle boundary discrepancies.

4.5 Automated QA Generation and Final Expert Validation

Using the comprehensive metadata assembled in previous steps (lesion presence, geometric information, and optimal/suboptimal masks), we automatically synthesize the full QA set for Stages 1–4. Every question and its corresponding answer is generated by a rule-based algorithm that maps the metadata into carefully designed textual templates, ensuring both scalability and linguistic consistency across the benchmark.

Recognizing that clinical validity is paramount, we conclude the pipeline with a rigorous expert review. The dataset is partitioned among a panel of six medical experts, and each expert independently inspects the visual prompts, option sets, and ground truth for their assigned items. In cases where the algorithmic output deviates from clinical judgment, the experts manually refine the labels. This final step guarantees that the benchmark strictly adheres to clinical standards.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and human baseline. We evaluate a total of 18 VLMs. These comprise 4 proprietary models (Gemini-3.1-pro, Gemini-3.1-flash (Team et al., 2023), GPT-5.4, and GPT-5.4-nano (Singh et al., 2025)). For open-source models, we select 5 from the general domain (Qwen3.6-27B (Qwen Team, 2026b), Qwen3.5-122B (Qwen Team, 2026a), GLM-4.6V (Hong et al., 2025), InternVL3.5-38B (Wang et al., 2025), Gemma4-31B (Team et al., 2024)) and 5 from the medical domain (MedGemma-27B (Sellergren et al., 2025), MedGemma1.5-4B (Sellergren et al., 2026), HuatuoGPT-Vision-7B (Chen et al., 2024a), Lingshu-32B (Xu et al., 2025), Hulu-Med-32B (Jiang et al., 2025)). The remaining 4 are the foundation backbones of the medical models, Gemma3-27B, Gemma3-4B, Qwen2.5-VL-32B, and Qwen2.5-VL-7B. We also report a human baseline, obtained by having three physicians each answer 24 questions, one per lesion and path combination, and averaging their scores.

Evaluation settings. We report results under two complementary settings. The *End-to-End* (E2E) setting is the strictest and most clinically realistic: if a model fails at any stage, every subsequent stage in the same sequence is also counted as incorrect, resulting in a strict per-stage accuracy. The *Oracle-Passed* (OP) setting is designed to isolate the model's upper-bound capability at each stage. Whenever the model answers incorrectly, we overwrite its previous answer in the conversation history with the ground-truth answer before proceeding, so that earlier errors do not penalize later stages, allowing us to reveal where perception truly breaks down.

Metrics. *Stage-level accuracy* (%) is the fraction of sequences in which the model successfully completes a given stage. For stages containing multiple sub-questions (Stages 3 and 4), the model must answer all sub-questions correctly to pass. *Depth* is defined as the average number of consecutive stages the model answers correctly before its first perceptual error under the E2E setting. To expose fine-grained perception weaknesses, we additionally report *sub-task accuracy* for Stages 3 and 4 under the OP setting.

5.2 RESULTS

Table 2 reports performance of 14 recent models under both settings, while Table 3 compares five medical VLMs against their backbones. Table 4 decomposes Stages 3 and 4 into their constituent sub-tasks. Below, we highlight several notable observations drawn from these results. Detailed performance breakdowns by lesion type and path, alongside additional are provided in Appendix E.

Stage-wise performance degradation. Under the E2E setting, all models exhibit a pronounced performance gap between Stage 1 and the subsequent stages. Although the strongest models at Stage 1 achieve above 80% accuracy, performance drops sharply at Stage 2 (a binary RR/RF routing task). Performance further collapses at Stage 3, where every open-source model scores below 2%; only Gemini-3.1-pro reaches 7.5%. Even at Stage 4, the top-performing GPT-5.4 resolves only 14.0% of all cases. In the OP setting, which explicitly blocks the propagation of prior mistakes, the average performance gains across models reveal that the marginal gain at Stage 3 ($+1.1$ pp) reflects a fundamental inability in contour revision, regardless of prior context. Conversely, while models also struggle intrinsically at Stage 2 and Stage 4, their comparatively larger gains ($+10.0$ pp and $+11.5$ pp) demonstrate that these stages are relatively more vulnerable to cascading upstream errors.

From the perspective of Depth, most models achieve above 0.7 on the LF path, whereas the RR and RF paths span roughly 1 to 2. However, the high scores in the RR and RF paths are largely artifacts of bias. Specifically, models that predominantly predict "No" (indicating no revision is needed) at Stage 2, such as Qwen3.6-27B, accumulate higher RF Depth at the expense of RR Depth. In contrast, MedGemma1.5-4B sits at the opposite extreme by strongly predicting revision necessity, achieving the highest RR Depth among all models. Detailed analyses of model-specific biases at Stage 2 are provided in Table 11 of Appendix E.4.

Physicians outperform every model by a wide margin at each stage. Their 59.5% at Stage 4 is lower than their scores elsewhere only because the stage requires all four attributes to be correct at once, and Table 4 shows that they answer every attribute but location (73.8%) with high accuracy.

Table 2: Stage-level accuracy (%) and per-path depth across recent 14 VLMs. Stage 1 is aggregated over all paths (RR, RF, LF), whereas Stages 2 and 4 are evaluated over {RR, RF}, and Stage 3 over RR only. Oracle-Passed Stage 1 is omitted because it equals End-to-End Stage 1 by construction. Cardiomegaly cases lack Stage 4; thus, clearing the last asked stage counts as passing the missing stage to keep the depth scale uniform. **Bold** values indicate the best performance within each group; **bold + underlined** values indicate the best across all groups. Green annotations in the Oracle-Passed section report percentage point gains over End-to-End performance. Detailed per-path accuracies are provided in Appendix Table 9.

Model	End-to-End				Oracle-Passed			Depth (E2E)		
	S1	S2	S3	S4	S2	S3	S4	RR (0-4)	RF (0-3)	LF (0-1)
Human	*97.2*	*91.7*	*70.8*	*50.0*	*95.8* (+4.2)	*79.2* (+8.3)	*59.5* (+9.5)	*3.08*	*2.50*	*1.00*
Random	*50.0*	*25.0*	*0.1*	*0.0*	*50.0* (+25.0)	*0.5* (+0.4)	*0.3* (+0.3)	*0.75*	*0.78*	*0.50*
GPT-5.4	82.3	44.6	2.2	**14.0**	55.5 (+10.9)	8.5 (+6.2)	35.4 (+21.4)	1.10	**1.76**	0.84
GPT-5.4-nano	71.5	34.4	0.0	2.5	52.2 (+17.8)	0.2 (+0.2)	13.0 (+10.5)	0.94	1.12	**0.84**
Gemini-3.1-pro	**86.9**	**50.4**	**7.5**	10.4	**56.6** (+6.2)	**10.0** (+2.5)	**38.9** (+28.5)	**1.53**	1.61	0.82
Gemini-3.1-flash	81.0	42.4	0.9	4.2	52.4 (+10.0)	1.1 (+0.2)	17.4 (+13.2)	1.21	1.42	0.82
Qwen3.6-27B	**89.7**	45.8	0.4	6.3	50.9 (+5.1)	**2.5** (+2.1)	13.7 (+7.4)	1.06	**1.88**	0.86
Qwen3.5-122B	86.2	40.2	0.0	**8.9**	50.0 (+9.8)	2.0 (+2.0)	**27.7** (+18.9)	1.07	1.61	**0.93**
GLM-4.6V	64.8	27.9	0.0	2.9	50.1 (+22.1)	1.1 (+1.1)	20.9 (+18.0)	0.56	1.27	0.83
InternVL3.5-38B	83.0	32.8	0.0	5.8	32.8 (+0.0)	0.1 (+0.1)	19.6 (+13.8)	0.93	1.63	0.78
Gemma4-31B	82.2	**46.7**	**1.5**	6.5	**54.2** (+7.6)	1.8 (+0.3)	24.6 (+18.1)	**1.47**	1.33	0.78
MedGemma-27B	72.8	32.9	0.0	1.6	**50.8** (+17.9)	**0.1** (+0.1)	3.4 (+1.9)	0.69	1.30	0.88
MedGemma1.5-4B	66.8	**49.8**	0.0	0.1	49.8 (+0.0)	0.0 (+0.0)	0.7 (+0.6)	**1.84**	1.17	0.00
HuatuoGPT-Vision-7B	62.2	27.9	0.0	0.3	49.6 (+21.8)	0.0 (+0.0)	1.6 (+1.4)	0.77	0.91	0.77
Lingshu-32B	**89.0**	43.3	0.0	**3.9**	47.9 (+4.6)	0.0 (+0.0)	**6.7** (+2.9)	0.97	**1.81**	0.89
Hulu-Med-32B	88.6	44.2	0.0	1.6	49.9 (+5.7)	0.0 (+0.0)	6.1 (+4.4)	1.36	1.36	**0.90**

Table 3: Comparison of open-source medical VLMs and their respective foundation backbone models on CheXpercept. Each medical-VLM cell is annotated with its change relative to the corresponding backbone cell: green (+) marks an improvement, red (−) marks a degradation, and black (+0.0) marks no change.

Model	End-to-End				Oracle-Passed			Depth (E2E)		
	S1	S2	S3	S4	S2	S3	S4	RR (0-4)	RF (0-3)	LF (0-1)
Gemma3-27B (Backbone)	57.6	28.5	0.0	1.4	52.9	0.0	7.0	0.69	0.95	0.69
MedGemma-27B	72.8 (+15.2)	32.9 (+4.4)	0.0 (+0.0)	1.6 (+0.2)	50.8 (−2.1)	0.1 (+0.1)	3.4 (−3.6)	0.69 (+0.00)	1.30 (+0.35)	0.88 (+0.19)
Gemma3-4B (Backbone)	61.2	37.6	0.0	0.0	50.6	0.0	0.1	1.29	0.96	0.38
MedGemma1.5-4B	66.8 (+5.6)	49.8 (+12.2)	0.0 (+0.0)	0.1 (+0.1)	49.8 (−0.8)	0.0 (+0.0)	0.7 (+0.6)	1.84 (+0.55)	1.17 (+0.21)	0.00 (−0.38)
Qwen2.5-VL-7B (Backbone)	52.8	21.6	0.0	0.1	50.8	0.0	2.1	0.44	0.98	0.72
HuatuoGPT-Vision-7B	62.2 (+9.4)	27.9 (+6.3)	0.0 (+0.0)	0.3 (+0.2)	49.6 (−1.2)	0.0 (+0.0)	1.6 (−0.5)	0.77 (+0.33)	0.91 (−0.07)	0.77 (+0.05)
Qwen2.5-VL-32B (Backbone)	66.5	31.0	0.0	1.9	48.7	0.0	14.6	1.11	0.91	0.63
Lingshu-32B	89.0 (+22.5)	43.3 (+12.3)	0.0 (+0.0)	3.9 (+2.0)	47.9 (−0.8)	0.0 (+0.0)	6.7 (−7.9)	0.97 (−0.14)	1.81 (+0.90)	0.89 (+0.26)
Hulu-Med-32B	88.6 (+22.1)	44.2 (+13.2)	0.0 (+0.0)	1.6 (−0.3)	49.9 (+1.2)	0.0 (+0.0)	6.1 (−8.5)	1.36 (+0.25)	1.36 (+0.45)	0.90 (+0.27)

Limitation of medical domain adaptation. Among open-source models, recent medical VLMs consistently match or underperform recent general VLMs of comparable or even smaller size across all stages. Lingshu-32B, the strongest medical VLM in Stage 1 (89.0%), collapses to 3.9% at Stage 4, falling behind not only larger general models but also the smaller Qwen3.6-27B. Hulu-Med-32B and MedGemma-27B likewise trail size-matched or smaller general counterparts (Gemma4-31B, Qwen3.6-27B). The gain from the OP setting also favors general models, most clearly at Stage 4: general models recover on average +15.2 pp once upstream errors are removed, whereas medical models gain only +2.2 pp.

To isolate the effect of domain adaptation itself, we further compare five medical VLMs against the general-domain backbones they were fine-tuned from (Table 3). Medical fine-tuning yields consistent gains on Stage 1, but the gains vanish or reverse on Stages 2, 3 and 4 under the OP setting. The Stage 2 gain under E2E is inherited from Stage 1 rather than earned. Consequently, the gains from current medical domain adaptation are limited in scope, and since recent general models outperform their medical counterparts at comparable sizes, the capability of the underlying backbone appears to matter more than the medical adaptation applied on top of it.

Table 4: Stage 3 and Stage 4 sub-task accuracy under the OP setting. Stage 3 evaluated on the RR path, comprises point-wise revision (Exp.: expansion, Con.: contraction) and revision-result selection (Res.). Stage 4 aggregated over the RR/RF paths, comprises semantic attributes: Dist. (distribution), Loc. (location), Sev. (severity), and Comp. (comparison). Highlighting follows Table 2.

Model	**Stage 3**			**Stage 4**			
	Exp.	Con.	Res.	Dist.	Loc.	Sev.	Comp.
Human	*87.5*	*87.5*	*91.7*	*92.9*	*73.8*	*97.6*	*92.9*
Random	*13.4*	*14.7*	*25.0*	*25.0*	*13.1*	*33.3*	*25.0*
GPT-5.4	**22.0**	41.1	**70.8**	70.5	**60.4**	80.8	81.9
GPT-5.4-nano	1.0	25.2	30.0	70.9	32.8	58.6	76.4
Gemini-3.1-pro	20.1	**55.2**	69.6	**76.3**	59.5	**81.9**	**88.8**
Gemini-3.1-flash	4.5	36.1	35.0	53.4	51.6	65.1	72.6
Qwen3.6-27B	**10.6**	25.5	31.2	51.5	48.1	39.1	57.0
Qwen3.5-122B	7.4	28.5	37.2	69.0	55.2	65.1	74.8
GLM-4.6V	8.4	**41.2**	32.9	72.1	39.2	67.1	81.3
InternVL3.5-38B	3.2	18.8	34.4	**72.4**	39.8	**70.4**	**82.7**
Gemma4-31B	4.9	40.1	**47.9**	53.5	**55.4**	56.5	51.0
MedGemma-27B	1.5	27.5	22.8	39.2	26.3	30.7	58.8
MedGemma1.5-4B	0.0	0.0	25.0	25.9	18.6	14.4	20.5
HuatuoGPT-Vision-7B	0.0	5.1	22.4	37.1	27.7	39.1	11.7
Lingshu-32B	0.0	**32.5**	19.6	41.9	35.1	41.4	**67.7**
Hulu-Med-32B	**4.9**	2.2	**25.2**	**54.5**	**39.8**	**43.6**	42.6

Sub-task analysis. At Stage 3, an asymmetry emerges where most models score substantially higher on contraction than expansion. We attribute this to a response bias. Models tend to predict points for expansion but default to "None" for contraction, which happens to be the correct answer in 46.6% of contraction cases (Table 6). A related bias appears in how many colored points a model proposes per case. Those that propose many (HuatuoGPT, MedGemma-27B, Lingshu-32B) score lower on point-wise metrics than those that propose few (Qwen3.6-27B, Hulu-Med-32B), so returning fewer points buys accuracy by accident rather than by perceptual gain. Per-model "None"-prediction rates and predicted-point counts are reported in Table 12 of Appendix E.5. Revision-result selection, a four-way visual comparison, is easier than point-wise reasoning for most models. GPT-5.4 leads at 70.8%, and Gemma4-31B is the strongest open-source model at 47.9%.

At Stage 4, proprietary and general open-source models consistently outperform medical models across all attributes. Gemini-3.1-pro takes the lead on three of the four attributes: distribution (76.3%), severity (81.9%), and comparison (88.8%), while GPT-5.4 leads on location (60.4%). Medical models consistently rank at the bottom regardless of attribute. Even the stronger medical models fall short of their general-purpose counterparts on every attribute: Hulu-Med-32B achieves 54.5% on distribution, 39.8% on location, 43.6% on severity, and Lingshu-32B reaches 67.7% on comparison, all below the corresponding scores of the top general and proprietary models. This pattern is consistent with the view that current medical adaptation may prioritize medical text patterns over the visual measurement and spatial localization skills required for attribute extraction. Physicians lead on every sub-task, though their score on location is somewhat lower, as it demands fine-grained anatomical grounding across 20 sub-regions.

6 Conclusion

In this study, we present **CheXpercept**, the first CXR benchmark that decomposes VLM perception into multiple sequential stages. It reveals that current VLMs fail at fine-level and semantic-level perception of major CXR lesions. Performance remains low even under the OP setting, which eliminates cascading errors from earlier stages, indicating that these perceptual limitations are intrinsic. We further find that recent medical VLMs are consistently worse than their general-purpose counterparts, and that adaptation from a general backbone improves only a limited range of perceptual levels, suggesting that current medical adaptation may focus more on medical text patterns than on strengthening visual perception of medical images. These findings underscore the need for new training paradigms that develop the sequential visual perception capabilities that radiologists employ.

AI USE STATEMENT

Generative AI tools were used in two capacities in this work: writing code that realizes the methods we designed, and correcting grammar to make the manuscript easier to read. Neither the research question, the approach we took to it, nor the experiments that evaluate it originated from such tools. Among the tasks for which disclosure is required, we did not use generative AI to synthesize data, to build theoretical models or conceptual frameworks, to state or prove mathematical claims, to form or sharpen hypotheses, to shape the design of our experiments, or to draw conclusions from the results. Nor did we rely on these tools to survey the literature or locate references. Every AI-assisted output passed through the authors: code was inspected, debugged, and tested for correctness before being relied upon, and each grammatical revision was read against the sentence it replaced to confirm that our scientific meaning survived the edit. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves the use of medical imaging data to evaluate the perceptual capabilities of vision-language models. To ensure adherence to ethical guidelines and patient privacy, we utilized strictly de-identified chest X-ray images sourced from publicly available datasets (MIMIC and SIIM-ACR).

To prevent the misuse of medical data and maintain rigorous data security, the proposed CheXpercept benchmark dataset will be hosted and distributed exclusively through PhysioNet. We will not release the raw data or annotations on unrestricted public repositories. Access will be strictly granted only to credentialed researchers who have completed mandatory human subjects research training (*e.g.*, CITI program) and signed a formal Data Use Agreement (DUA).

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our work, we provide detailed information regarding both the dataset construction and the evaluation framework. First, the detailed protocols for data generation, including the medical experts' review process, are extensively described in the Appendix. The proposed CheXpercept benchmark dataset will be made available via PhysioNet upon publication, requiring standard credentialing for access.

REFERENCES

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, et al. Mimic-ext-mimic-cxr-vqa: a complex, diverse, and large-scale visual question answering dataset for chest x-ray images. *PhysioNet*, 2024.

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

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.

Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna. Radiotransformer: A cascaded global-focal transformer for visual attention–guided disease classification. In *European Conference on Computer Vision*, pp. 679–698. Springer, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024a.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b.

Geon Choi, Hangyul Yoon, Hyunju Shin, Hyunki Park, Sang Hoon Seo, Eunho Yang, and Edward Choi. Instruction-guided lesion segmentation for chest x-rays with automatically generated large-scale dataset. *arXiv preprint arXiv:2511.15186*, 2025.

Geon Choi, Hangyul Yoon, Hyunju Shin, Hyunki Park, Sang Hoon Seo, Eunho Yang, and Edward Choi. MIMIC-CXR-Ext-ILS: Lesion Segmentation Masks and Instruction-Answer Pairs for Chest X-rays. *PhysioNet*, March 2026. doi: 10.13026/8ejy-4t06. URL https://doi.org/10.13026/8ejy-4t06. Version 1.0.0.

Matias Cosarinsky, Nicolas Gaggion, Rodrigo Echeveste, and Enzo Ferrante. Chexmask-u: Quantifying uncertainty in landmark-based anatomical segmentation for x-ray images. *arXiv preprint arXiv:2512.10715*, 2025.

Andrew J Degnan, Emily H Ghobadi, Peter Hardy, Elizabeth Krupinski, Elena P Scali, Lindsay Stratchko, Adam Ulano, Eric Walker, Ashish P Wasnik, and William F Auffermann. Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. *Academic radiology*, 26(6):833–845, 2019.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4156–4165, 2023.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical sam3: A foundation model for universal prompt-driven medical image segmentation. *arXiv preprint arXiv:2601.10880*, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels. *PhysioNet*, March 2024a. doi: 10.13026/jsn5-t979. URL https://doi.org/10.13026/jsn5-t979. Version 2.1.0.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database. *PhysioNet10*, 13026(C2JT1Q):5, 2024b.

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

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

Hyungyung Lee, Geon Choi, Jung-Oh Lee, Hangyul Yoon, Hyuk Gi Hong, and Edward Choi. Cxreasonbench: A benchmark for evaluating structured diagnostic reasoning in chest x-rays. *arXiv preprint arXiv:2505.18087*, 2025.

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

Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21310–21320, 2025.

Jong Hak Moon, Geon Choi, Paloma Rabaey, Min Gwan Kim, Hyuk Gi Hong, Jung-Oh Lee, Hangyul Yoon, Eun Woo Doe, Jiyoun Kim, Harshita Sharma, et al. Lunguage: A benchmark for structured and sequential chest x-ray interpretation. *arXiv preprint arXiv:2505.21190*, 2025.

Andriy Myronenko, Dong Yang, Baris Turkbey, Mariam Aboian, Sena Azamat, Esra Akcicek, Hongxu Yin, Pavlo Molchanov, Marc Edgar, Yufan He, et al. Reasoning visual language model for chest x-ray analysis. *arXiv preprint arXiv:2510.23968*, 2025.

Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pp. 251–264. World Scientific, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

Constantin Seibold, Alexander Jaus, Matthias A Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling. *arXiv preprint arXiv:2306.03934*, 2023.

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

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Shaoxuan Wu, Jingkun Chen, Chong Ma, Cong Shen, Xiao Zhang, and Jun Feng. Following the diagnostic trace: Visual cognition-guided cooperative network for chest x-ray diagnosis. *arXiv preprint arXiv:2602.21657*, 2026.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. SIIM-ACR Pneumothorax Segmentation, 2019. URL `https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation`.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A BENCHMARK STATISTICS

A.1 LESION TYPES

The scope of our benchmark covers a substantial portion of critical CXR findings. An LLM-based frequency analysis of 218K MIMIC-CXR reports shows that our selected types account for approximately 72% of all reported findings, covering all major lesions with a prevalence over 3% (effusion 17.2%, pneumothorax 13.8%, opacity 8.5%, atelectasis 8.1%, edema 6.8%, cardiomegaly 6.6%, consolidation 6.4%, pneumonia 4.4%).

A.2 PER-LESION COMPOSITION

Table 5 reports the per-lesion composition of CheXpercept. Each lesion contributes 100 evaluation sequences per path (RR, RF, LF), yielding 300 CXRs per lesion. Cardiomegaly produces fewer QA items per sequence than other lesions because Stage 4 attribute extraction is not applicable.

Table 5: Statistics of CheXpercept.

Lesion	Sequences per Path			QA Items per Sequence			CXRs	QA Items
	RR	RF	LF	RR	RF	LF		
Cardiomegaly	100	100	100	5	2	1	300	800
Pneumonia	100	100	100	9	6	1	300	1,600
Pneumothorax	100	100	100	9	6	1	300	1,600
Atelectasis	100	100	100	9	6	1	300	1,600
Opacity	100	100	100	9	6	1	300	1,600
Consolidation	100	100	100	9	6	1	300	1,600
Edema	100	100	100	9	6	1	300	1,600
Effusion	100	100	100	9	6	1	300	1,600
Total	700	700	700				2,400	12,000

A.3 GROUND-TRUTH ANSWER DISTRIBUTION

Table 6 reports the ground-truth answer distribution for every sub-question in CheXpercept. Stages 1 and 2 are constructed so that the three paths carry an equal number of cases. The low rate of "L larger" in the Stage 4 comparison task arises because the heart anatomically overlaps the left lung, effectively reducing the left-lung area available for lesion masks.

Table 6: Ground-truth answer distribution of CheXpercept by stage and sub-task. For Stage 3, the distribution compares "None" (no points required) versus "Specific" (one or more colored points required, with the mean point count shown in parentheses). Stage 4 excludes cardiomegaly cases. Dashes indicate options that are not applicable to the corresponding sub-task.

Sub-task	**Path**	**N**	**Gold answer distribution**			
			Option 1	**Option 2**	**Option 3**	**Option 4**
Stage 1: Detection						
Detection	RR	800	Yes (100%)	No (0%)	–	–
Detection	RF	800	Yes (100%)	No (0%)	–	–
Detection	LF	800	Yes (0%)	No (100%)	–	–
Stage 2: Contour Evaluation						
Contour Eval	RR	800	Yes (100%)	No (0%)	–	–
Contour Eval	RF	800	Yes (0%)	No (100%)	–	–
Stage 3: Contour Revision						
Expansion	RR	800	None (31.2%)	Specific (68.8%, mean 1.29)	–	–
Contraction	RR	800	None (46.6%)	Specific (53.4%, mean 1.12)	–	–
Revision-Result	RR	800	Opt 1 (23.9%)	Opt 2 (28.2%)	Opt 3 (26.4%)	Opt 4 (21.5%)
Stage 4: Attribute Extraction (non-cardiomegaly)						
Distribution	RR/RF	1,400	1 zone (39.6%)	2 zones (27.1%)	3 zones (9.8%)	$\geq$4 zones (23.4%)
Location	RR/RF	1,400	1 region (26.6%)	2 regions (25.9%)	3 regions (21.9%)	none of the above (25.6%)
Severity	RR/RF	1,400	$<$1/3 lung (54.6%)	1/3–2/3 lung (32.7%)	$\geq$2/3 lung (12.7%)	–
Comparison	RR/RF	1,400	single lung (70.6%)	similar L/R (14.6%)	R larger (12.4%)	L larger (2.4%)

B BENCHMARK CONSTRUCTION DETAILS

B.1 EXTERNAL DATASETS AND MODELS

The CheXpercept construction pipeline builds on the following publicly available datasets and pretrained models.

MIMIC-CXR-JPG. MIMIC-CXR-JPG (Johnson et al., 2024a;b; 2019) is a large-scale publicly available chest X-ray dataset consisting of CXRs and associated radiology reports collected at the Beth Israel Deaconess Medical Center in Boston, MA. It serves as the original resource for the MIMIC-ILS dataset (Choi et al., 2025; 2026).

- URL: https://physionet.org/content/mimic-cxr-jpg/2.1.0/
- License: PhysioNet Credentialed Health Data License 1.5.0

MIMIC-ILS. MIMIC-ILS (Choi et al., 2025; 2026) is a large-scale CXR lesion segmentation dataset derived from MIMIC-CXR-JPG (Johnson et al., 2024a;b; 2019). It covers seven key lesion types (cardiomegaly, pneumonia, atelectasis, opacity, consolidation, edema, effusion), pairing each CXR with textual instructions (*e.g.*, "Segment the pneumonia."), target-lesion presence labels, and lesion masks produced by an automated pipeline optimized for clinically acceptable quality. We use its training split as the source of candidate CXRs and lesion masks for both the abnormal and normal pools.

- URL: https://physionet.org/content/mimic-cxr-ext-ils/1.0.0/
- License: PhysioNet Credentialed Health Data License 1.5.0

SIIM-ACR Pneumothorax Segmentation. SIIM-ACR Pneumothorax Segmentation (Zawacki et al., 2019) is a public CXR dataset released by the Society for Imaging Informatics in Medicine (SIIM) and the American College of Radiology (ACR) for a segmentation challenge. It contains both positive and negative cases, where each positive case is accompanied by a radiologist-annotated mask.

- URL: https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation/overview
- License: Kaggle competition rules (academic and commercial use permitted)

ROSALIA. ROSALIA (Choi et al., 2025) is a vision-language model fine-tuned on MIMIC-ILS for prompt-driven CXR lesion segmentation. Given a CXR and a textual instruction specifying a target lesion, it outputs a binary mask and a text description. We leverage ROSALIA to re-infer cleaner lesion masks across the MIMIC-ILS training split prior to expert review (Appendix B.2).

- URL: https://github.com/checkoneee/ROSALIA
- License: Apache-2.0 license

CheXmask-U. CheXmask-U (Cosarinsky et al., 2025) is a landmark-based anatomical segmentation dataset providing high-quality masks for the left and right lungs. We use the HybridGNet model, pretrained on CheXmask-U, as one of our sources for lung masks.

- Dataset URL: https://huggingface.co/datasets/mcosarinsky/CheXmask-U
- Model URL: https://github.com/mcosarinsky/CheXmask-U
- Dataset License: Apache license 2.0
- Model License: GPL-3.0 license

CXAS. CXAS (Seibold et al., 2023) is an anatomy segmentation model capable of delineating 159 chest anatomical structures in CXRs. We extract only its lung masks and fuse them with the HybridGNet outputs through a preprocessing step (Appendix B.3). The resulting consolidated lung masks serve as the geometric foundation for our 20-region partitioning algorithm (Appendix B.4).

- URL: https://physionet.org/content/chexmask-cxr-segmentation-data/
- License: CC BY 4.0

SAM3. SAM3 (Carion et al., 2025) is a promptable foundation segmentation model that accepts point, box, and mask prompts to delineate target objects. We repurpose it as a precise mask deformer: using an expert-curated optimal mask as the initial mask prompt, alongside automatically sampled positive and negative point prompts in disjoint lung sub-regions, SAM3 generates the suboptimal masks required for Stages 2 and 3 (Appendix B.5).

We use the original SAM3 rather than a medical variant such as Medical SAM3 (Jiang et al., 2026), for three reasons. First, our pipeline manipulates masks through point prompts, whereas the medical variants are optimized for boxes or text; in our tests they responded poorly to points, which makes them unsuitable here. Second, they are fine-tuned on large, well-defined anatomical structures, a bias that offers little for the irregular and sparse morphologies of the eight lesions in our benchmark. Third, and most importantly, we do not ask SAM3 to segment an image from scratch. It acts purely as a deformation engine applying controlled perturbations to masks that are already clinically optimal, and what matters for that role is how faithfully the model follows the geometric intent of a prompt, not the domain it was trained on.

- URL: https://github.com/facebookresearch/sam3
- License: SAM License (https://github.com/facebookresearch/sam3/blob/main/LICENSE)

B.2 Mask refinement via ROSALIA re-inference

The lesion masks distributed in MIMIC-ILS are generated by an automated pipeline; consequently, a subset exhibits noisy artifacts, such as jagged boundaries or small disconnected components. To establish a cleaner candidate pool for expert review, we re-infer all masks across the training split using ROSALIA (Choi et al., 2025). Although ROSALIA was trained on these same source masks, its predictions exhibit greater spatial consistency and smoother boundaries. We therefore replace the original annotations with these re-inferred versions prior to expert curation (§4.1). Importantly, this refinement serves solely as a preprocessing cleanup step: the final optimal masks in CheXpercept are strictly those that subsequently pass rigorous expert review.

B.3 Lung mask preparation

Before the 20-region partitioning (Appendix B.4) and downstream suboptimal mask generation (Appendix B.5), we preprocess the lung masks so that they are anatomically tight yet still cover the entire lesion. Two pretrained anatomy segmentation models, HybridGNet (Cosarinsky et al., 2025) (trained on CheXmask-U) and CXAS (Seibold et al., 2023), are first applied to each CXR to obtain two independent estimates of the left and right lungs. We intersect the two estimates per side, yielding strict lung masks M_L^0 and M_R^0 that are robust to the systematic biases of either model alone.

When the target lesion (*e.g.*, consolidation) extends slightly beyond the strict lung boundary, the strict lung masks may inadvertently truncate the lesion during region partitioning. To prevent this, we expand the lung masks using the optimal lesion mask. Concretely, each connected component of the optimal lesion mask is unioned into the left mask if it overlaps M_L^0, and into the right mask if it overlaps M_R^0. The resulting masks, M_L and M_R, serve as the inputs for Algorithm 1 in Appendix B.4.

B.4 20-region lung partitioning algorithm

The 20 sub-regions used throughout CheXpercept (for lesion-location ground truth in Stage 4 and for disjoint prompt sampling in suboptimal mask generation, §4.4) are obtained by intersecting two complementary partitions of the prepared lung masks M_L, M_R from Appendix B.3. Algorithm 1 formalizes the procedure; we describe each component below.

Vertical zones. Each lung is divided vertically into *upper*, *middle*, and *lower* zones by splitting its vertical extent into three equal bands.

Horizontal sets. Each lung is split horizontally at its vertical midline into a *medial* half (closer to the body midline) and a *lateral* half. Additionally, a *peripheral* set is derived by morphologically eroding the bilateral lung union with a square footprint—whose size is proportional to the local lung width—and subsequently subtracting the resulting central core. This peripheral set effectively captures the outer rim of the lungs.

Region categories. The 18 zone-aligned sub-regions (9 per side) are obtained through the pairwise intersections of the three horizontal sets $\{\text{medial}, \text{lateral}, \text{peripheral}\}$ and the three vertical zones $\{\text{upper}, \text{middle}, \text{lower}\}$. Furthermore, we delineate the *costophrenic angle* (the bottom quarter of the peripheral set) as a distinct zone for each side due to its clinical significance. The complete 20-region partition is visualized in Figure 2. All ratios utilized in this procedure (*e.g.*, the $1/3$ vertical splits and the $1/4$ costophrenic-angle cutoff) were established through consultations with medical experts.

Algorithm 1 20-region lung partitioning.

Require: Left lung mask M_L, right lung mask M_R
Ensure: Set $\mathcal{R}$ of 20 sub-region masks

$\mathcal{R} \leftarrow \emptyset$
for all $s \in \{L, R\}$ **do** ▷ Step 1: vertical thirds (zones)
  $(y^s_{\min}, y^s_{\max}) \leftarrow$ vertical extent of M_s; $\quad h_s \leftarrow y^s_{\max} - y^s_{\min} + 1$
  $b_1 \leftarrow y^s_{\min} + \lfloor h_s/3 \rfloor$; $\quad b_2 \leftarrow y^s_{\min} + \lfloor 2h_s/3 \rfloor$
  $Z^{\text{up}}_s \leftarrow M_s \cap \{y \le b_1\}$; $\quad Z^{\text{mid}}_s \leftarrow M_s \cap \{b_1 < y \le b_2\}$; $\quad Z^{\text{low}}_s \leftarrow M_s \cap \{y > b_2\}$
end for
for all $s \in \{L, R\}$ **do** ▷ Step 2: medial / lateral split
  $x^s_{\text{mid}} \leftarrow$ midpoint of horizontal extent of M_s
  $A^{\text{med}}_s \leftarrow$ portion of M_s on the body-midline side of x^s_{mid}
  $A^{\text{lat}}_s \leftarrow M_s \setminus A^{\text{med}}_s$
end for
$w \leftarrow$ mean lung width of M_R at the $1/3$ and $2/3$ vertical positions ▷ Step 3: peripheral set
$K \leftarrow$ square footprint of side $\lfloor w/2 \rfloor$; $\quad C \leftarrow \text{erode}(M_L \cup M_R,\, K)$
for all $s \in \{L, R\}$ **do**
  $P_s \leftarrow M_s \setminus C$ ▷ outer rim of lung s
end for
for all $s \in \{L, R\}$ **do** ▷ Step 4: zone-aligned sub-regions (3×3 per lung)
  for all $H \in \{A^{\text{med}}_s, A^{\text{lat}}_s, P_s\}$ **do**
    for all $Z \in \{Z^{\text{up}}_s, Z^{\text{mid}}_s, Z^{\text{low}}_s\}$ **do**
      $\mathcal{R} \leftarrow \mathcal{R} \cup \{H \cap Z\}$
    end for
  end for
end for
for all $s \in \{L, R\}$ **do** ▷ Step 5: costophrenic angle (bottom $1/4$ of peripheral)
  $y^s_{\text{cut}} \leftarrow y^s_{\min} + \lfloor 3h_s/4 \rfloor$
  $\text{CPA}_s \leftarrow P_s \cap \{y > y^s_{\text{cut}}\}$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{\text{CPA}_s\}$
end for
return $\mathcal{R}$ ▷ $|\mathcal{R}| = 9 \times 2 + 2 = 20$

B.5 SAM3-BASED DEFORMATION MECHANICS

This subsection details how point prompts are sampled and assembled to drive the SAM3-based mask deformer described in §4.4. A single concrete pass of the full procedure is illustrated in Figure 4.

Sub-region assignment. For each case, we first group the 20 sub-regions into six broad anatomical zones: the upper, middle, and lower zones for both the left and right lungs (where the costophrenic angle is assigned to the lower zone). Within each zone, we randomly select one sub-region overlapping with the optimal mask and assign it a deformation operation (expansion or contraction).

Let r_{sub} denote the fraction of the selected sub-region covered by the lesion, and r_{lung} denote the fraction of the corresponding lung covered by the lesion. To ensure the deformation remains anatomi-

cally plausible, this random operation assignment is overridden under the following conditions: (1) contraction is forced if $r_{\text{sub}} > 0.75$ (as the sub-region is already nearly filled, leaving little room for expansion); (2) expansion is forced if $r_{\text{sub}} < 0.5$ (as too little lesion is present in the sub-region for contraction to be meaningful); and (3) contraction is forced whenever $r_{\text{lung}} > 0.7$ (as the lesion already saturates the lung).

The resulting (sub-region, operation) pairs collectively form the complete deformation plan for the case. From this plan, we randomly sample a subset of pairs to determine the specific locations where actual point prompts will be fed into SAM3. For instance, in Figure 4, the sampled subset includes an expansion operation for the medial portion of the left mid zone and a contraction operation for the left costophrenic angle. For cardiomegaly, where the target lies inside a single anatomical structure rather than spanning multiple lung sub-regions, this same logic is applied directly at the zone level, bypassing the medial/lateral/peripheral subdivision.

Direction by polarity. Each point prompt is generated by sampling from the specific segment of the transformed optimal mask's contour that overlaps with the previously assigned sub-region. For *expansion*, we dilate the optimal mask with an elliptical kernel and extract its contour, which lies just outside the original lesion boundary. The portion of this contour that falls within the assigned sub-region forms the pool of candidate *positive* prompts (label 1), driving SAM3 to expand the mask outward. For *contraction*, we instead erode the optimal mask and extract its contour, which lies just inside the original boundary. The portion of this contour that falls within the assigned sub-region forms the pool of candidate *negative* prompts (label 0), driving SAM3 to contract the mask inward. In the third panel of Figure 4, the yellow bands trace the dilation rings used for expansion candidates, and the green and purple bands trace the erosion rings used for contraction candidates; the green stars and red crosses represent the actual positive and negative prompts sampled from those overlapping segments, respectively.

Magnitude by depth and width. The magnitude of each deformation is controlled by two integer parameters per sub-region. *Depth* d specifies the total number of concentric contour rings from which prompts are sampled. Specifically, at each depth level $i \in \{1, \ldots, d\}$, the optimal mask is dilated (or eroded) by $i \times t$ morphological iterations, where t is a fixed step size. The contour of the mask at each level is extracted, resulting in d distinct contours. With $d = 2$, for example, points are sampled from both the contour at one dilation step ($i = 1$) and the contour at two steps ($i = 2$), jointly anchoring the deformation across multiple offsets from the original boundary.

Width determines the number of points placed on each contour, with a minimum spacing of 30 pixels between them. The width at depth 1 is randomly sampled from $\{2, 3, 4\}$ and is constrained to be non-increasing across subsequent depths. Furthermore, points at depth $i > 1$ are positioned near the points at depth $i - 1$, ensuring that successive depths extend the existing prompt cluster rather than initiating new, isolated ones.

Throughout CheXpercept, we fix $d = 2$, meaning each deformation inherently consists of two layers of prompts. The example in Figure 4 uses widths $[D_1{:}W_3,\ D_2{:}W_2]$ for expansion and $[D_1{:}W_4,\ D_2{:}W_2]$ for contraction. For visual clarity, the main-paper illustration (Figure 3) simplifies this two-layer sampling to a single contour and represents the magnitude axis solely through width, which it refers to as the *deformation level*.

Disjoint sampling. When a single optimal mask is subjected to both expansion and contraction in a single pass, the two prompt sets must not compete for the same local area; otherwise, SAM3's boundary update becomes unstable. While the deformation plan inherently restricts expansion and contraction operations to distinct sub-regions, opposing prompts sampled near a shared boundary could still cause conflicts. Therefore, we additionally enforce a minimum distance of 50 pixels between any expansion-contraction prompt pair. Figure 4 satisfies this constraint: the expansion prompts in the medial portion of the left mid zone and the contraction prompts at the left costophrenic angle lie in non-adjacent sub-regions, safely exceeding the distance threshold. Once the prompts are collected from their respective sub-regions, they are concatenated and passed to SAM3 alongside the initial optimal mask. SAM3 then generates the final suboptimal mask in a single forward pass (final panel of Figure 4).

Sequential application for QA. For each item, we randomly select 1 to 3 sub-regions from the deformation plan and apply the deformations step-by-step (expansions first, followed by contractions), saving an intermediate suboptimal mask after each step. This yields a sequence of suboptimal masks per case that share the same lesion identity but differ in the cumulative number of corrupted sub-regions. In parallel, after each step, we generate a set of *distractor* masks rooted in the same intermediate state. Concretely, we re-invoke SAM3 using prompts with incorrect polarity or targeting the wrong sub-regions. These distractor prompts are sampled from morphologically transformed copies of the intermediate mask, following the same dilation/erosion-and-contour procedure used for the true deformation. Consequently, the resulting distractor masks shift the boundary in an incorrect direction, yet retain strong geometric similarity to the correctly revised mask due to their shared intermediate origin. The true revised mask and its distractors together form the option set for the Stage 3 multiple-choice question, forcing the model to discriminate the correct revision from plausible same-step alternatives that share the overall lesion geometry.

Handling fallback distractors. Occasionally, generating a sufficiently diverse set of distractors fails, for instance when the original lesion is exceptionally small or when internal deformation constraints (*e.g.*, sub-region restrictions) limit valid prompt placements. If the distractor count falls short, we pad the remaining options with predefined default masks, such as the unrevised suboptimal mask itself or the full left or right lung mask. These fallbacks account for 12.9% of all options.

Quality of suboptimal masks. The suboptimal masks differ from their optimal counterparts by a clear margin, with a mean IoU of 0.715 (standard deviation 0.154). That the difference is also visible to the eye is borne out by the human baseline, whose accuracy on Stage 2 in Table 2 is high.

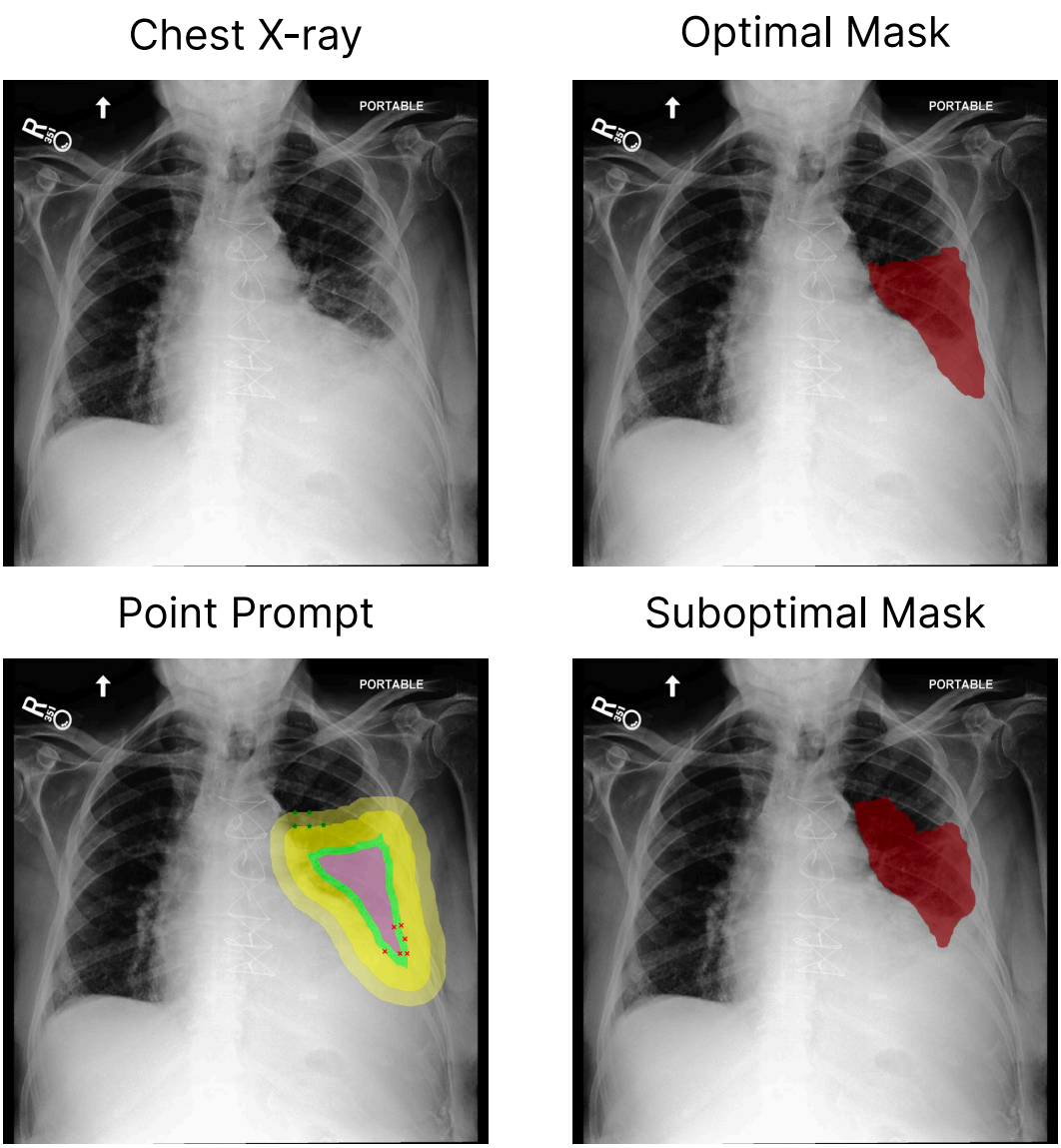


Figure 4: End-to-end illustration of a single deformation step on a pneumonia case in which one expansion sub-region and one contraction sub-region are applied simultaneously to the same lesion component. The expansion is configured with Depth d=2 and Widths $[D_1{:}W_3,\ D_2{:}W_2]$ (5 positive prompts in total, shown as green stars) and is anchored at the medial portion of the *left mid zone*. The contraction uses d=2 and Widths $[D_1{:}W_4,\ D_2{:}W_2]$ (6 negative prompts, shown as red crosses) at the *left costophrenic angle*. Panels (left to right): the original CXR; the optimal mask overlaid on the CXR; the dilation (yellow) and erosion (green and purple) contour rings together with the sampled point prompts; the SAM3 input (optimal mask combined with both prompt sets); and the resulting suboptimal mask returned by SAM3 in a single forward pass.

C QA ITEM SPECIFICATIONS AND PROMPT TEMPLATES

This appendix documents the exact prompt templates, option strings, and answer encoding used to generate each CheXpercept QA item. The placeholder {`lesion_name`} is substituted with one of the eight target lesions (atelectasis, cardiomegaly, consolidation, edema, effusion, opacity, pneumonia, pneumothorax).

System prompt. At evaluation time, every model receives a fixed system message (Figure 5) that frames it as an expert radiologist and constrains its response to a single line of the form "Answer: [Option Number]", with comma-separated indices permitted for multi-select questions.

> You are an expert radiologist. Analyze the given chest X-ray image and the question. Provide the final answer in the following format: 'Answer': [Option Number] Only the option number(s) should be included in the final answer part. (e.g., 'Answer: 1', 'Answer: 2', 'Answer: 1,3').

Figure 5: System message used for all evaluations.

C.1 STAGE 1: LESION PRESENCE DETECTION

This foundational question is asked across all evaluation paths. The model is provided with the original chest X-ray and is required to output a binary response, encoded as `answer_index` $\in \{1, 2\}$ (where 1 corresponds to "Yes" and 2 to "No"). We employ two question stem variants (Figure 6): a default formulation and a specialized finding-style formulation for opacity and consolidation, which mirrors the way these findings are articulated in standard radiology reports. Both prompt formats present straightforward Yes/No options.

> **Default form (used for atelectasis, cardiomegaly, edema, effusion, pneumonia, pneumothorax)**
>
> Is any finding suggestive of {lesion_name} visible in the image?
>
> Options:
> (1) Yes
> (2) No
>
> **Form used for opacity and consolidation**
>
> Is any {lesion_name} visible in the image?
>
> Options:
> (1) Yes
> (2) No

Figure 6: Stage 1 (detection) prompt templates.

C.2 STAGE 2: CONTOUR EVALUATION

This stage is evaluated along both the RR (suboptimal mask) and RF (optimal mask) paths. The model is provided with a chest X-ray in which the candidate lesion mask is rendered as a colored overlay. Specifically, the model is asked to determine whether the presented mask requires major revision (Figure 7), responding with a binary choice encoded as `answer_index` $\in \{1, 2\}$ (1 for "Yes", 2 for "No"). Additionally, for non-cardiomegaly lesions, we append a specific constraint instructing the model to ignore any lesions that overlap with the heart.

C.3 STAGE 3: CONTOUR REVISION

Conducted exclusively on the RR path, this stage is structured as a multi-turn conversation encompassing three sub-tasks (Figure 8, illustrated with a four-point example): point-wise expansion (3a), point-wise contraction (3b), and revision-result selection (3c). Stages 3a and 3b present up to eight

The image shows a chest X-ray with a predicted mask for the {lesion_name} (shown in the overlay).
Is any major revision needed?
Note that lesions overlapping with the heart do not need to be considered.

Options:
(1) Yes
(2) No

Figure 7: Stage 2 (contour evaluation) prompt template.

Initial context

The colored points indicate potential locations where the mask may need to be adjusted. The mask may need to be expanded towards these points, contracted at these locations, or both. Note that lesions overlapping with the heart do not need to be considered.

Stage 3a - Point-wise expansion

Which point(s), if any, should be used to expand the mask toward areas of the {lesion_name} that are not currently covered? Select all points where expansion is needed. If no expansion is needed, select 'None'.

Options:
(1) red point
(2) orange point
(3) yellow point
(4) green point
(5) None

Stage 3b - Point-wise contraction

Which point(s), if any, should be used to contract the mask away from areas that are not part of the {lesion_name}? Select all points where contraction is needed. If no contraction is needed, select 'None'.

Options:
(1) red point
(2) orange point
(3) yellow point
(4) green point
(5) None

Stage 3c - Revision-result selection

Choose the mask that reflects the modifications.

Options:
(1) 1st mask
(2) 2nd mask
(3) 3rd mask
(4) 4th mask

Figure 8: Stage 3 (contour revision) prompt templates.

colored point markers over the mask overlay and ask the model to identify which points should be used to expand or contract the mask. Consistent with other stages, for non-cardiomegaly cases, the model is explicitly instructed to ignore any areas overlapping with the heart. The expected response is a list of the chosen option numbers. Finally, Stage 3c involves a four-way visual comparison among the true revised mask and three distractors, which are generated from the same intermediate state under different deformation patterns (Appendix B.5).

C.4 STAGE 4: ATTRIBUTE EXTRACTION

Conducted on the RR and RF paths exclusively for non-cardiomegaly cases, this stage asks the model for the geometric properties a physician reads off a lesion once it is delineated (Figure 9). The evaluation is divided into four attribute-focused sub-tasks: (4a) *distribution*, which queries the number of pulmonary zones the lesion occupies; (4b) *location*, a multi-select query covering three anatomical sub-regions from the 20-region partition (Appendix B.4) plus a "None of the above" option; (4c) *severity*, assessing the proportion of a selected lung occupied by the lesion; and (4d) *comparison*, a bilateral size assessment in which "larger" is strictly defined as having an area $\geq 1.5\times$ that of the contralateral lesion. For all relevant sub-tasks, lung areas explicitly exclude the heart silhouette.

Stage 4a - Distribution

In how many zones is the {lesion_name} present in the image? Consider the following zones: right upper zone lung, right mid zone lung, right lower zone lung, left upper zone lung, left mid zone lung, left lower zone lung.

Options:
(1) 1 zone
(2) 2 zones
(3) 3 zones
(4) 4 zones or more

Stage 4b - Location

Where is the {lesion_name} located in the image? Select all locations where the lesion is present from the options.

Options:
(1) {anatomy_option_1}
(2) {anatomy_option_2}
(3) {anatomy_option_3}
(4) None of the above (options 1-3)

Stage 4c - Severity

What proportion of the {selected_lung} area does the {lesion_name} occupy? Note that the lung area excludes the heart area.

Options:
(1) Less than 1/3 of the {selected_lung} area
(2) 1/3 or more but less than 2/3 of the {selected_lung} area
(3) 2/3 or more of the {selected_lung} area

Stage 4d - Comparison

Assess and compare the sizes of the lesion across both lungs. If the lesion is present in only one lung, select the corresponding option. Note that the lung area excludes the heart area. A lesion is considered larger if it is 1.5 times or more the size of the other.

Options:
(1) Both lesions are similar in size.
(2) The left lung lesion is larger than the right lung lesion.
(3) The right lung lesion is larger than the left lung lesion.
(4) The lesion is present in only one lung (left or right).

Figure 9: Stage 4 (attribute extraction) prompt templates.

D QUALITATIVE EXAMPLES

This appendix walks through complete CheXpercept items as they are presented to the model, illustrating the visual content, options, and ground-truth answers at each stage. Options highlighted in green denote the gold answers. The text-only prompt templates that wrap these visuals are given in Appendix C.

Consolidation (LF). Figure 10 illustrates a lesion-free (LF) case for consolidation. Because the CXR is a true-normal image with no visible consolidation, the ground truth for Stage 1 is "No". For all LF items, the evaluation pipeline terminates immediately after this initial detection step; consequently, no candidate masks are presented, and neither the contour evaluation nor the attribute extraction stages are conducted.

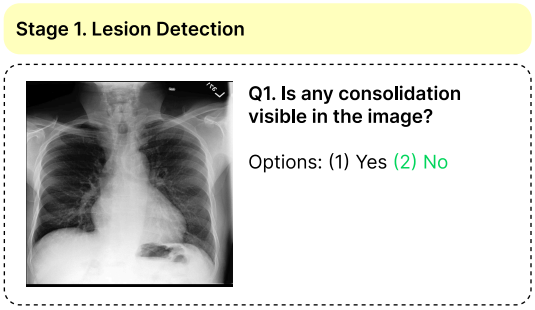


Figure 10: Evaluation pipeline for a lesion-free (LF) consolidation case.

Effusion (RF). Figure 11 illustrates a revision-free (RF) case for effusion. After Stage 1 confirms the lesion's presence, Stage 2 determines that the provided optimal mask does not require major revision. Consequently, the Stage 3 contour revision step is skipped, and the pipeline proceeds directly to Stage 4, where four sub-questions probe the semantic attributes of the lesion.

Cardiomegaly (RR). Figure 12 details a revision-required (RR) case for cardiomegaly. Following presence confirmation in Stage 1, Stage 2 identifies the deformed candidate mask as requiring revision. Stage 3 then presents three sequential sub-tasks on the same image: an expansion query, a contraction query, and a four-way revision-result selection. By design, the evaluation for cardiomegaly terminates after Stage 3, since the four Stage 4 attributes (distribution, location, severity, comparison) are defined exclusively for focal lung lesions.

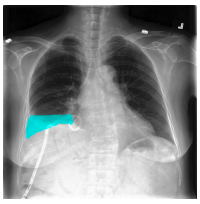
Stage 1. Lesion Detection



Q1. Is any finding suggestive of effusion visible in the image?

Options: (1) Yes (2) No

Stage 2. Lesion Contour Evaluation



Q2. The image shows a chest X-ray with a predicted mask for the effusion (shown in the overlay). Is any major revision needed? Note that lesions overlapping with the heart do not need to be considered.

Options: (1) Yes (2) No

Stage 4. Lesion Attribute Extraction

Q4-1 (Distribution). In how many zones is the effusion present in the image? Consider the following zones: right upper zone lung, right mid zone, right lower zone lung, left upper zone lung, left mid zone lung, left lower zone lung.

Options:
(1) 1 zone
(2) 2 zones
(3) 3 zones
(4) 4 zones or more

Q4-2 (Location). Where is the effusion located in the image? Select all locations where the lesion is present from the options.

Options:
(1) the right costophrenic angle
(2) the intersection of the right peripheral lung and the right lower zone lung
(3) the intersection of the right peripheral lung and the right upper zone lung
(4) None of the above (options 1-3)

Q4-3 (Severity). What proportion of the right lung area does the effusion occupy? Note that the lung area excludes the heart area.

Options:
(1) Less than 1/3 of the right lung area
(2) 1/3 or more but less than 2/3 of the right lung area
(3) 2/3 or more of the right lung area

Q4-4 (Comparison). Assess and compare the sizes of the lesion across both lungs. If the lesion is present in only one lung, select the corresponding option. Note that the lung area excludes the heart area. A lesion is considered larger if it is 1.5 times or more the size of the other.

Options:
(1) Both lesions are similar in size.
(2) The left lung lesion is larger than the right lung lesion.
(3) The right lung lesion is larger than the left lung lesion.
(4) The lesion is present in only one lung (left or right).

Figure 11: Evaluation pipeline for an effusion revision-free (RF) case. Because the optimal mask is judged as not requiring revision at Stage 2, Stage 3 is skipped, and the item proceeds directly to the Stage 4 attribute extraction sub-tasks. Ground-truth (gold) answers are highlighted in green.

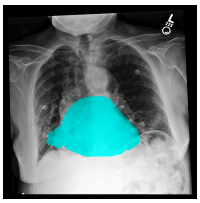
Stage 1. Lesion Detection



Q1. Is any finding suggestive of cardiomegaly visible in the image?

Options: (1) Yes (2) No

Stage 2. Lesion Contour Evaluation

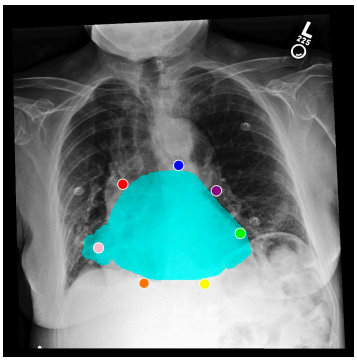


Q2. The image shows a chest X-ray with a predicted mask for the cardiomegaly (shown in the overlay). Is any major revision needed?

Options: (1) Yes (2) No

Stage 3. Lesion Contour Revision

The colored points indicate potential locations where the mask may need to be adjusted. The mask may need to be expanded towards these points, contracted at these locations, or both. Note that lesions overlapping with the heart do not need to be considered.

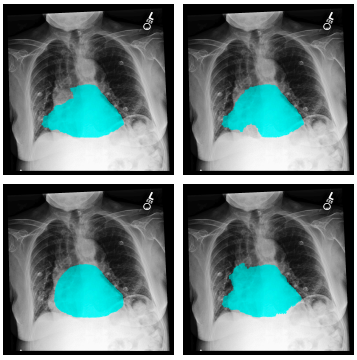


Q3-1. Which point(s), if any, should be used to expand the mask toward areas of the cardiomegaly that are not currently covered? Select all points where expansion is needed. If no expansion is needed, select 'None'.

(The options are the same as Q3-2. The answer is None)

Q3-2. Which point(s), if any, should be used to contract the mask away from areas that are not part of the cardiomegaly? Select all points where contraction is needed. If no contraction is needed, select 'None'.

Options:
(1) red point
⋮
(7) pink point
(8) None



Q3-3: Choose the mask that reflects the modifications.

Options:
(1) 1st mask
(2) 2nd mask
(3) 3rd mask
(4) 4th mask

Figure 12: Evaluation pipeline for a cardiomegaly revision-required (RR) case. Ground-truth answers are highlighted in green. Note that the evaluation concludes at Stage 3, as the Stage 4 attribute extraction is not applicable to cardiomegaly.

E EVALUATION SETUP AND DETAILED ANALYSIS

E.1 EVALUATED MODELS AND INFERENCE CONFIGURATION

Model deployment and privacy compliance. Table 7 summarizes the 18 vision-language models benchmarked in this paper. Because CheXpercept is built on the credentialed MIMIC-CXR-JPG dataset, we strictly adhere to the PhysioNet guidelines regarding the responsible use of such data with online services. Specifically, the PhysioNet data use agreement (DUA) prohibits sharing data through standard public APIs. To comply with these requirements, open-source models were downloaded from the Hugging Face Hub and evaluated locally on a cluster of $8\times$ NVIDIA A100 SXM4 (80 GB) GPUs. We utilized tensor parallelism for acceleration; however, a few models were run with a lower tensor-parallel degree due to architectural constraints, such as attention-head or expert counts not being divisible by eight. Furthermore, all proprietary models were accessed exclusively through specialized enterprise configurations that satisfy PhysioNet's criteria for data privacy: GPT-5.4 and GPT-5.4-nano via the Azure OpenAI Service, and Gemini-3.1-pro and Gemini-3.1-flash via Google Cloud Vertex AI.

Sampling and inference configuration. All open-source models are served locally using the vLLM engine with greedy decoding (`temperature=0.0`, `top_p=1.0`, `max_tokens=8192`), with the exception of Hulu-Med-32B which is served via the Hugging Face `transformers` library. For Qwen3.6-27B and Qwen3.5-122B, the deep-thinking mode is disabled to maintain manageable latency at the benchmark's scale. Proprietary models are queried via their respective enterprise APIs at the default temperature (1.0) with an output budget of 8,192 tokens per turn. For these models, the `reasoning_effort` (GPT-5.4 series) and `thinking_level` (Gemini-3.1 series) parameters are consistently set to `medium`.

Inference cost report. The total expenditure for proprietary models (Table 8) is approximately \$274. GPT-5.4 ($\sim$\$132) and Gemini-3.1-pro ($\sim$\$117) account for the bulk of this cost, while the lightweight nano and flash variants together represent less than 10% of the total. Both proprietary and open-source models completed the 2,400-case evaluation within approximately 24 to 72 hours of wall-clock time.

Table 7: Details of the 18 primary vision-language models evaluated on CheXpercept.

Display name	Params	Version / Hugging Face ID	Backend
Proprietary			
GPT-5.4	–	`gpt-5.4-2026-03-05`	Azure OpenAI API
GPT-5.4-nano	–	`gpt-5.4-nano-2026-03-17`	Azure OpenAI API
Gemini-3.1-pro	–	`gemini-3.1-pro-preview`	Vertex AI API
Gemini-3.1-flash	–	`gemini-3.1-flash-lite`	Vertex AI API
General (open-source)			
Qwen3.6-27B	27B	`Qwen/Qwen3.6-27B`	vLLM
Qwen3.5-122B	122B	`Qwen/Qwen3.5-122B-A10B`	vLLM
GLM-4.6V	106B	`zai-org/GLM-4.6V`	vLLM
InternVL3.5-38B	38B	`OpenGVLab/InternVL3_5-38B`	vLLM
Gemma4-31B	31B	`google/gemma-4-31B-it`	vLLM
Medical (open-source)			
MedGemma-27B	27B	`google/medgemma-27b-it`	vLLM
MedGemma1.5-4B	4B	`google/medgemma-1.5-4b-it`	vLLM
HuatuoGPT-Vision-7B	7B	`FreedomIntelligence/HuatuoGPT-Vision-7B-Qwen2.5VL`	vLLM
Lingshu-32B	32B	`lingshu-medical-mllm/Lingshu-32B`	vLLM
Hulu-Med-32B	32B	`ZJU-AI4H/Hulu-Med-32B`	`transformers`
Foundation backbones (open-source)			
Gemma3-27B	27B	`google/gemma-3-27b-it`	vLLM
Gemma3-4B	4B	`google/gemma-3-4b-it`	vLLM
Qwen2.5-VL-32B	32B	`Qwen/Qwen2.5-VL-32B-Instruct`	vLLM
Qwen2.5-VL-7B	7B	`Qwen/Qwen2.5-VL-7B-Instruct`	vLLM

Table 8: Approximate per-model API spend on the full CheXpercept benchmark (2,400 cases, 12,000 API calls each). Values are rounded from per-case `token_usage` log entries; provider billing may differ slightly.

Model	Input tok.	Cached (%)	Output tok.	Cost (USD)
GPT-5.4	∼34M	∼52	∼6M	∼$132
GPT-5.4-nano	∼34M	∼41	∼3M	∼ $ 8
Gemini-3.1-pro	∼50M	∼34	∼4M	∼$117
Gemini-3.1-flash	∼50M	∼25	∼5M	∼ $17
Total	∼167M	∼36	∼18M	∼$274

E.2 PER-PATH STAGE ACCURACY

The main-paper Table 2 reports per-stage accuracy averaged across the three paths. Table 9 provides the underlying per-path breakdown for every model under both End-to-End (E2E) and Oracle-Passed (OP) settings.

Table 9: Per-path stage-level accuracy (%) for each evaluated model. Path codes follow the main paper: RR (revision-required), RF (revision-free), LF (lesion-free). Stage 1 is identical between E2E and OP by construction, Dashes denote stages that do not exist for that path. Averaging the rows of this table across paths reproduces the per-stage accuracies in Table 2.

Model	**Path**	**S1**	**End-to-End**			**Oracle-Passed**		
			S2	S3	S4	S2	S3	S4
Proprietary								
GPT-5.4	RR	82.6	24.0	2.2	1.1	31.6	8.5	33.4
	RF	80.2	65.2	–	26.9	79.4	–	37.4
	LF	84.0	–	–	–	–	–	–
GPT-5.4-nano	RR	64.4	29.2	0.0	0.0	54.4	0.2	10.6
	RF	66.1	39.5	–	5.0	50.0	–	15.4
	LF	84.1	–	–	–	–	–	–
Gemini-3.1-pro	RR	91.1	51.6	7.5	2.7	55.2	10.0	32.0
	RF	87.4	49.1	–	18.1	58.0	–	45.9
	LF	82.1	–	–	–	–	–	–
Gemini-3.1-flash	RR	79.9	39.8	0.9	0.0	51.2	1.1	14.1
	RF	81.5	45.1	–	8.4	53.6	–	20.7
	LF	81.6	–	–	–	–	–	–
General (open-source)								
Qwen3.6-27B	RR	92.5	12.8	0.4	0.0	15.0	2.5	13.4
	RF	90.6	78.8	–	12.6	86.8	–	14.0
	LF	85.9	–	–	–	–	–	–
Qwen3.5-122B	RR	84.1	22.8	0.0	0.0	31.4	2.0	24.4
	RF	81.6	57.6	–	17.7	68.6	–	31.0
	LF	92.8	–	–	–	–	–	–
GLM-4.6V	RR	56.0	0.0	0.0	0.0	0.1	1.1	18.4
	RF	55.9	55.9	–	5.7	100.0	–	23.3
	LF	82.6	–	–	–	–	–	–
InternVL3.5-38B	RR	85.4	7.2	0.0	0.0	7.2	0.1	13.6
	RF	85.6	58.4	–	11.6	58.4	–	25.6
	LF	78.0	–	–	–	–	–	–
Gemma4-31B	RR	85.5	59.9	1.5	0.3	69.8	1.8	21.3
	RF	83.5	33.5	–	12.7	38.8	–	27.9
	LF	77.5	–	–	–	–	–	–
Medical (open-source)								
MedGemma-27B	RR	65.8	3.5	0.0	0.0	16.1	0.1	3.3
	RF	64.6	62.3	–	3.1	85.4	–	3.6
	LF	87.9	–	–	–	–	–	–
MedGemma1.5-4B	RR	100.0	84.5	0.0	0.0	84.5	0.0	0.4
	RF	100.0	15.1	–	0.1	15.1	–	1.0
	LF	0.4	–	–	–	–	–	–
HuatuoGPT-Vision-7B	RR	54.9	21.8	0.0	0.0	49.4	0.0	1.0
	RF	54.9	34.0	–	0.6	49.9	–	2.3
	LF	76.9	–	–	–	–	–	–
Lingshu-32B	RR	89.0	8.1	0.0	0.0	10.4	0.0	4.3
	RF	89.0	78.5	–	7.7	85.4	–	9.1
	LF	89.1	–	–	–	–	–	–
Hulu-Med-32B	RR	89.8	46.1	0.0	0.0	52.4	0.0	3.7
	RF	86.2	42.2	–	3.3	47.4	–	8.4
	LF	89.8	–	–	–	–	–	–

E.3 PER-LESION STAGE ACCURACY

While main-paper results aggregate performance across all eight target lesions, we provide a more granular view in Table 10 by reporting per-lesion stage accuracy for Gemini-3.1-pro, our strongest baseline. We focus on this single model to ensure that performance variances across lesions are attributable to lesion-specific characteristics rather than model-level heterogeneity. Note that, unlike the main-paper Depth (Table 2), the per-lesion Depth here is reported on each lesion's native scale: for cardiomegaly, the maxima are 3 (RR) and 2 (RF) because Stage 4 is undefined, whereas the main-paper Depth treats clearing the last asked stage of cardiomegaly as also clearing the missing Stage 4 to keep RR/RF maxima uniform across lesions (4 and 3, respectively).

Table 10: Per-lesion stage-level accuracy (%) and per-path Depth for Gemini-3.1-pro. Each lesion contributes 300 sequences (100 per path: RR, RF, LF). Stage 1 is identical between E2E and OP settings by construction and is reported once. Depth is the average number of consecutive correct stages under E2E setting, with maxima of 4, 3, and 1 for RR, RF, and LF respectively (for cardiomegaly, 3, 2, and 1 since Stage 4 does not apply). Dashes denote stages that do not apply for the lesion.

Lesion	**S1**	**End-to-End**			**Oracle-Passed**			**Depth (E2E)**		
		S2	S3	S4	S2	S3	S4	RR (0-4)	RF (0-3)	LF (0-1)
Atelectasis	86.7	58.5	11.0	14.0	58.5	13.0	36.0	1.75	1.78	0.63
Cardiomegaly	80.7	43.5	2.0	–	60.0	4.0	–	1.02	1.35	0.94
Consolidation	96.3	48.0	5.0	10.5	50.0	14.0	34.5	1.39	1.78	0.94
Edema	95.7	56.5	3.0	12.0	57.0	4.0	25.0	1.39	1.95	0.93
Effusion	93.7	53.5	28.0	11.0	56.5	28.0	55.5	2.30	1.21	0.87
Opacity	84.0	57.0	5.0	8.0	57.5	5.0	27.5	1.74	1.60	0.53
Pneumonia	92.7	57.5	3.0	14.5	60.0	8.0	33.5	1.49	1.90	0.86
Pneumothorax	65.3	28.5	3.0	3.0	53.5	4.0	60.5	1.13	0.62	0.87

E.4 STAGE 2 RESPONSE BIAS

Table 11 provides empirical evidence for the response bias discussed in the main text. To quantify this bias at Stage 2, we calculate the disparity in each model's "Yes" response rate between the RR and RF paths: $\Delta = \text{Rate}_{\text{Yes, RR}} - \text{Rate}_{\text{Yes, RF}}$. For nine out of the ten open-source VLMs, the absolute difference is negligible ($|\Delta| < 5$ pp), indicating that their near-50% Stage 2 accuracy is not a result of a partial perceptual signal, but rather an arithmetic consequence of a fixed class prior. Among the open-source models, only Gemma4-31B exhibits a non-trivial gap of $+8.5$ pp (accuracy 54.2%). The proprietary Gemini-3.1-pro shows the largest overall disparity ($+13.2$ pp, accuracy 56.6%).

E.5 STAGE 3 SUB-TASK BIAS

Table 12 provides the per-model evidence for the two response biases identified at Stage 3 in the main text. Most models predict "None" for contraction at much higher rates than for expansion (*e.g.*, Gemini-3.1-pro: 60.1% vs 1.9%, Gemma4-31B: 66.6% vs 3.2%, GLM-4.6V: 84.5% vs 5.9%), which directly produces the contraction $>$ expansion accuracy asymmetry. Across all fourteen models, the average number of predicted points per case has a strong negative correlation with accuracy: Pearson $r = -0.68$ for expansion and $r = -0.77$ for contraction, supporting the main text's claim that proposing many points incidentally lowers point-wise accuracy.

Table 11: Stage 2 (contour evaluation) per-model response distribution. RR (gold = yes) is the deformed-mask path; RF (gold = no) is the optimal-mask path. Yes/No/Etc (%) are the rates at which the model answers "1" (revision needed), "2" (no revision), or anything else (e.g., textual replies such as "Answer: No"). Δ is the discrimination gap (yes-rate on RR minus yes-rate on RF); a model with no perceptual signal has $\Delta \approx 0$.

Model	**RR**			**RF**			Δ	**Acc.**
	Yes	No	Etc	Yes	No	Etc		
Proprietary								
GPT-5.4	31.6	68.4	0.0	20.6	79.4	0.0	+11.0	55.5
GPT-5.4-nano	54.4	45.6	0.0	50.0	50.0	0.0	+4.4	52.2
Gemini-3.1-pro	55.2	44.8	0.0	42.0	58.0	0.0	+13.2	56.6
Gemini-3.1-flash	51.2	48.8	0.0	46.4	53.6	0.0	+4.9	52.4
General (open-source)								
Qwen3.6-27B	15.0	85.0	0.0	13.2	86.8	0.0	+1.8	50.9
Qwen3.5-122B	31.4	68.6	0.0	31.4	68.6	0.0	+0.0	50.0
GLM-4.6V	0.1	99.9	0.0	0.0	100.0	0.0	+0.1	50.1
InternVL3.5-38B	7.2	58.5	34.2	8.4	58.4	33.2	-1.1	32.8
Gemma4-31B	69.8	30.2	0.0	61.2	38.8	0.0	+8.5	54.2
Medical (open-source)								
MedGemma-27B	16.1	83.9	0.0	14.5	85.4	0.1	+1.6	50.8
MedGemma1.5-4B	84.5	15.5	0.0	84.9	15.1	0.0	-0.4	49.8
HuatuoGPT-Vision-7B	49.4	50.6	0.0	50.1	49.9	0.0	-0.8	49.6
Lingshu-32B	10.4	89.6	0.0	14.6	85.4	0.0	-4.2	47.9
Hulu-Med-32B	52.4	47.6	0.0	52.6	47.4	0.0	-0.2	49.9

Table 12: Stage 3 (contour revision) per-model response bias and accuracy on revision-required cases. "None" (%) is the rate at which the model select the number for "None" (no point selected), "Pts" is the mean number of predicted points per case, and "Acc." is exact-match accuracy against the gold point set. From Table 6, the ground-truth "None" rates are 31.2% (expansion) and 46.6% (contraction), so a model that always answers "None" would already match the gold for that fraction of cases.

Model	**Expansion**			**Contraction**		
	None (%)	Pts	Acc. (%)	None (%)	Pts	Acc. (%)
Proprietary						
GPT-5.4	17.4	1.60	22.0	32.8	1.16	41.1
GPT-5.4-nano	0.4	4.39	1.0	31.4	1.15	25.2
Gemini-3.1-pro	1.9	2.52	20.1	60.1	0.63	55.2
Gemini-3.1-flash	1.8	3.34	4.5	64.4	0.88	36.1
General (open-source)						
Qwen3.6-27B	3.6	2.03	10.6	45.9	1.16	25.5
Qwen3.5-122B	1.8	2.40	7.4	50.8	0.83	28.5
GLM-4.6V	5.9	3.97	8.4	84.5	0.20	41.2
InternVL3.5-38B	2.8	3.26	3.2	31.8	1.32	18.8
Gemma4-31B	3.2	3.30	4.9	66.6	0.70	40.1
Medical (open-source)						
MedGemma-27B	3.2	7.75	1.5	61.9	2.29	27.5
MedGemma1.5-4B	0.0	5.90	0.0	0.0	5.93	0.0
HuatuoGPT-Vision-7B	0.0	5.17	0.0	9.0	4.19	5.1
Lingshu-32B	0.0	5.33	0.0	67.9	0.49	32.5
Hulu-Med-32B	0.0	2.21	4.9	1.8	1.93	2.2

F EXPERT ANNOTATION

This appendix documents the manual annotation procedures within the CheXpercept pipeline. Sections F.1 and F.2 describe the two candidate-pool curation protocols (optimal-mask and true-normal CXR filtering); Section F.3 details the final QA validation step; and Section F.5 provides the professional profiles of the six medical experts who carried out all of the above tasks (cf. §4).

F.1 OPTIMAL MASK FILTERING

Each candidate item in the abnormal pool is uniquely identified by a `key_id` of the form `{lesion}_{study_id}_positive_{index}`. For every candidate, we generate a single annotation panel that places the original CXR side-by-side with the same image overlaid by the candidate mask in red, together with three header fields drawn from the MIMIC-ILS metadata: `key_id` (item identifier), `mapped_location` (the lesion location parsed from the original radiology report), and `segmentation_source` (the instruction type under which ROSALIA produced the candidate mask; the value "`global`" indicates that the mask was inferred from the lesion-only instruction "Segment the {*lesion_name*}." rather than from a location-conditioned prompt). Figure 13 shows a representative panel.

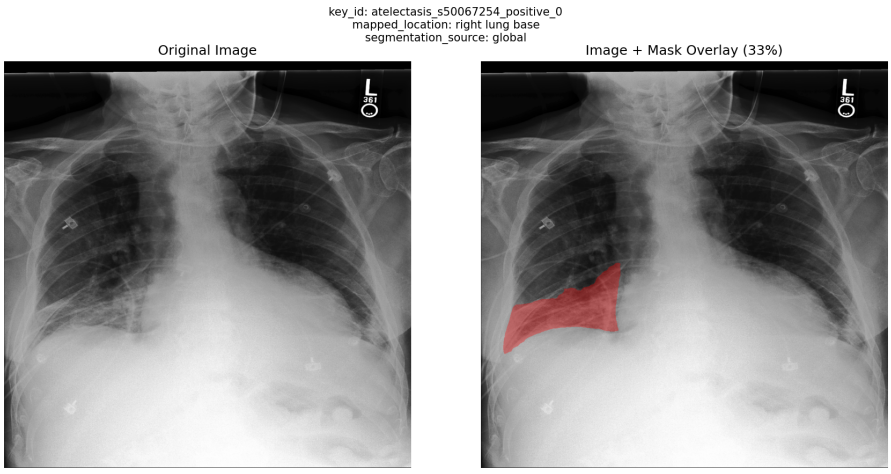


Figure 13: Annotation panel shown to experts during optimal-mask filtering.

Experts log their decisions on a shared spreadsheet (Figure 14). Each row carries the responsible `annotator`, the lesion `target`, and a single `optimal?` cell that is marked "`O`" when the candidate mask precisely traces the lesion boundary and left blank otherwise. Only candidates with a positive "`optimal?`" mark enter the optimal-mask pool; all others are discarded.

F.2 TRUE-NORMAL CXR FILTERING

For LF path candidates, experts review the raw CXR alone (without any mask overlay) and decide whether the image is genuinely free of a target lesion. Decisions are logged on a sheet analogous to the one used in §F.1, with a single "`normal?`" column marked "`O`" for accepted true-normal cases and left blank for rejected ones. Only "`O`"-marked candidates enter the true-normal pool that supplies LF path items.

F.3 FINAL QA VALIDATION

Following the automated QA generation (§4.5), every item in the assembled dataset undergoes a rigorous manual verification process to ensure absolute clinical accuracy. Each QA item is rendered into a comprehensive visualization—combining the input image, question stem, options, and algorithmically assigned ground truth—which is then independently vetted by our board-certified medical experts. The experts perform an item-by-item inspection, logging corrections on a sheet.

optimal_mask_filtering

File Edit View Insert Format Data Tools Gemini Extensions Help

	annotator	key_id	target	optimal?
2	D	atelectasis_s50009350_positive_0	atelectasis	O
3	D	atelectasis_s50045407_positive_1	atelectasis	
4	D	atelectasis_s50067254_positive_0	atelectasis	O
5	D	atelectasis_s50081893_positive_0	atelectasis	O
6	D	atelectasis_s50178702_positive_2	atelectasis	
7	D	atelectasis_s50183991_positive_2	atelectasis	
8	D	atelectasis_s50247129_positive_0	atelectasis	
9	D	atelectasis_s50290018_positive_0	atelectasis	O
10	D	atelectasis_s50402322_positive_2	atelectasis	
11	D	atelectasis_s50436993_positive_1	atelectasis	O
12	D	atelectasis_s50500733_positive_1	atelectasis	O
13	D	atelectasis_s50518588_positive_0	atelectasis	
14	D	atelectasis_s50579159_positive_6	atelectasis	O
15	D	atelectasis_s50644361_positive_0	atelectasis	
16	D	atelectasis_s50647734_positive_0	atelectasis	
17	D	atelectasis_s50755798_positive_0	atelectasis	O
18	D	atelectasis_s50773973_positive_0	atelectasis	
19	D	atelectasis_s50847846_positive_0	atelectasis	
20	D	atelectasis_s50866509_positive_2	atelectasis	O

Figure 14: Decision sheet used during optimal-mask filtering.

In any instance where the algorithmic output deviates from expert clinical judgment, the expert's decision overrides the initial label.

F.4 Inter-Annotator Agreement

In our two-round human annotation workflow consisting of initial optimal mask filtering and a final QA evaluation, cases were randomly redistributed during the final QA evaluation. This resulted in 288 masks being re-evaluated by the same annotator and 1,207 masks being cross-evaluated by a different annotator. Among the 1,207 cross-evaluated masks, 1,146 (94.9%) were evaluated as optimal again. Because the pipeline pre-filters cases marked optimal in the first round, the initial label variance is zero, making Cohen's kappa degenerate. Therefore, we report Gwet's AC1 (Gwet, 2008), which is robust to marginal imbalance. The inter-annotator AC1 was 0.946 (95% CI: 0.932 to 0.960), indicating almost perfect agreement.

To measure IAA on the final QA fix annotations, reviewers independently re-annotated a subset of cases completed by another annotator. Across the pooled 1,360 attribute-sub-question judgments and 140 contour-revision judgments, Gwet's AC1 was 0.959 (95% CI: 0.947–0.969) for attribute extraction and 0.971 (95% CI: 0.941–0.993) for contour revision, both in the "almost perfect" range per Landis and Koch. We believe these results, together with our initial optimal mask filtering agreement (AC1 = 0.946, 95% CI: 0.932–0.960), demonstrate that our two-round human annotation and verification workflow maintains almost perfect consistency across all stages and all lesion types.

F.5 Medical Expert Profiles

The benchmark construction and final QA validation were carried out by a panel of six medical experts (anonymized as A–F for double-blind review). Experts A–F are all radiation oncologists with 6, 4, 4, 11, 8, and 8 years of clinical experience in lesion segmentation, respectively.