npj Digital Medicine

https://doi.org/10.1038/s41746-025-02292-9

Article in Press

Text-image alignment for ILD imaging: linking CXR evidence to CT quantification

Received: 2 November 2025

Accepted: 16 December 2025

Cite this article as: Gao, J., Ren, Y., Yang, F. *et al.* Text-image alignment for ILD imaging: linking CXR evidence to CT quantification. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-025-02292-9

Jiani Gao, Yijiu Ren, Fengjing Yang, Xuefei Hu, Changbo Sun, Sihua Wang & Chang Chen

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Text–Image Alignment for ILD Imaging: Linking CXR Evidence to CT Quantification

Jiani Gao[1†], Yijiu Ren[1†], Fengjing Yang[2†], Xuefei Hu[1*], Changbo Sun[1*], Sihua Wang[2*], Chang Chen[1,3*]

[1]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, 200433, Shanghai, People's Republic of China.
[2]Department of Thoracic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430022, Hubei, People's Republic of China.
[3]Clinical Center for Thoracic Surgery Research, Tongji University, Shanghai, 200433, Shanghai, People's Republic of China.

*Corresponding author(s). E-mail(s): huxuefei_12345@163.com; changbosun@tongji.edu.cn; sihua_wang@hust.edu.cn; chenthoracic@163.com;
Contributing authors: 18217793797@163.com; ryjscott@hotmail.com; m202376172@hust.edu.cn;
[†]These authors contributed equally to this work.

Abstract

Assessment of Interstitial Lung Disease (ILD) relies on chest radiographs (CXR) for screening and computed tomography (CT) for definitive quantification. However, current AI pipelines typically treat these modalities in isolation, leading to report hallucinations and cross-modal inconsistencies. To address this fragmentation, we propose a framework (ARCTIC-ILD) that aligns CXR-derived textual evidence with CT-level segmentation and quantification. The system first employs a calibrated CXR evidence extractor to map radiographs to ILD-specific terminology, producing structured findings. These findings condition a terminology-to-mask module that utilizes lightweight cross-attention adapters to generate lobe-aware CT masks and burden estimates. Crucially, an explicit vision–language audit enforces consistency between the generated text and quantitative data. Evaluations on paired CXR–CT cohorts demonstrate that the framework significantly reduces text hallucination and improves phrase-to-mask alignment without incurring additional inference latency. By coupling reporting with quantification under an auditable protocol, this approach aligns with clinical workflows, serving as a robust assistant for triage, structured reporting, and longitudinal follow-up.

1 Introduction

Interstitial lung disease (ILD) is an important cause of chronic respiratory failure and death, and fibrosis represents the central pathologic manifestation as well as a key marker of disease progression [1, 2]. In routine care pathways, chest radiography (CXR) is frequently obtained earlier and more often because of its accessibility for initial screening and longitudinal follow up, whereas thin section high resolution CT (HRCT) is the reference standard for detailed depiction and quantitative assessment of parenchymal involvement [3]. Characteristic fibrosis related signs relevant to ILD include reticulation, honeycombing, and traction bronchiectasis, for which standardized definitions are provided by professional society glossaries and clinical practice guidelines [4]. Clinical decision making subsequently relies on HRCT to characterize the spatial distribution and burden of disease at lobar and related anatomic scales, which enables reproducible reporting and supports quantitative analyses [2, 5]. Within this context, a unified computational framework that systematically couples the high accessibility and early cueing afforded by CXR with the fine grained quantification available from HRCT, while auditing the consistency between generated textual reports and structured quantitative outputs, addresses core requirements for traceable and clinically actionable assistance in ILD care [5, 6].

In recent years, vision-language models (VLMs) and large language models (LLMs) have demonstrated substantial potential for radiology report generation and multimodal understanding [7–9]. For chest radiography, the prevailing paradigm has progressed from coupling an image encoder with an autoregressive text decoder to frameworks that foreground cross-modal alignment, instruction tuned generation, and the integration of clinically oriented evaluability [6–8]. Domain specific pretraining that explicitly leverages temporal structure has provided a more stable foundation for vision-language representations in radiology and helps mitigate semantic drift associated with learning solely from static paired data [7]. At the label level, large scale chest radiograph corpora such as CheXbert define a reusable set of fourteen observations with explicit handling of uncertainty and an accompanying evaluation protocol, thereby furnishing standardized probabilistic outputs that can serve as upstream conditions for evidence constrained generation [3, 10].

Concurrently, advances in zero-shot discriminative objectives indicate that self-supervised or contrastive image-text matching can generalize to previously unseen terminology without explicit labels, offering a practical mechanism for regularizing phrasing consistency outside the training domain and for providing a reference for confidence estimation [11]. On the generator side, multi stage training that combines connector alignment between vision and language spaces with instruction tuning followed by small step convergence, together with data organization that couples image conditioning and radiology writing conventions, has yielded increasingly operational report pipelines and initiated expert blinded assessments and evaluations in realistic settings focused on clinical factuality [5, 6, 8, 9]. Meanwhile, the community has acknowledged that surface level automatic text metrics are limited. To address the gap where high lexical overlap coexists with clinically significant errors, it has introduced entity level and relation level measures (such as RadGraph F1 and the composite score RadCliQ) along with broader imaging aligned evaluation frameworks [5, 6].

In text driven medical segmentation, open-vocabulary vision advances provide tools for phrase to mask inference under weak or zero-shot supervision. One line uses promptable unified decoders supporting multiple prompts for downstream tasks [12]. Another leverages CLIP based dense prediction via lightweight decoders, enabling segmentation without closed label sets [13, 14]. In medical CT, recent work explores terminology driven pretraining, contrastive objectives, and cross domain transfer for text controllable segmentation [15]. Mature open lung and lobar segmentation toolchains enable cross case comparison [16, 17]. With spatial functions, subpleural fibrosis is quantified into reproducible indices [5, 12]. Emerging benchmarks for grounding text free radiology to volumetric masks motivate sentence level supervision for phrase guided 3D segmentation [13, 14].

Notwithstanding this progress, three limitations recur. First, controllability is limited. Multimodal report generators produce free form text without hard constraints from downstream evidence, leading to prompt drift, evidence drift, and factual inconsistencies [5, 6]. Second, a semantic spatial disconnect persists. CT spatial qualifiers are not routinely tied to verifiable voxel level evidence, despite enabling consensus terminology [13, 14]. Third, auditing and reproducibility are under specified, and modern CT methods still face phrase ambiguity, inter slice discontinuity, and incomplete anatomic normalization [12, 15]. These indicate a need for multimodal methodology driven by controllable evidence chains, constrained by auditable structured outputs, and supported by reproducible procedures.

We propose ARCTIC-ILD, an Audited, Reproducible, and Controllable Vision–Language Coupling for Interstitial Lung Disease. Orchestrated by a multimodal agent backbone, it comprises two tightly coupled branches: CXR evidence and controlled reporting. For report generation, we use BioViL-T (backbone unchanged) with a Multi-Label Evidence Head (EVH) added to its image branch [7]. EVH is supervised on CheXbert's 14 observations to yield probability vectors for four key ILD findings as explicit evidence [10]. EVH outputs are calibrated via Temperature Calibrating with fixed thresholds (CAL) to align with generation controls [18]. BioViL-T visual features are projected into visual tokens, which with a control prefix from high confidence EVH outputs feed an instruction tuned text generator [8, 9]. Training follows LLaVA's recipe to constrain generation to visual evidence [8, 19]. Decoding ties target tokens to region representations with attention biased by calibrated probabilities to reduce factual drift. Inspired by CheXzero [11], an ILD specific contrastive image–text matching head regularizes cross-modal learning in training and yields sentence level consistency scores at inference. The agent ultimately generates auditable radiology paragraphs via a stable interface [11]. CT terminology to mask with audited quantification. For text-guided CT segmentation, we build terminology driven segmentation or quantification to map clinical language to imaging findings, addressing three slice wise challenges: coarse boundaries, poor inter slice coherence, and error cascading. We use text prompt driven SEEM with low rank adapters (LoRA) in cross-modal attention for lightweight ILD adaptation [12, 19]. To train the Terminology2Mask Module (TRM), we pair phrases from the Fleischner glossary derived phrase bank with HUG-ILD annotations for terminology to mask supervision. We also employ hard negative contrastive learning to enhance the discriminability of terms [4, 13, 14]. We introduce Text-Conditioned Diffusion Refinement (TCDR) to refine boundaries and generate uncertainty maps via few step diffusion sampling, and a training free trellis module, named Streaming Memory (STM) to improve cross slice coherence and suppress errors. At inference, integrated outputs with lobar segmentation yield key quantifications and structured descriptions, providing auditable CT semantic parsing [5, 16].

Finaly, VILA-M3 serves as the top level agent that orchestrates the workflow through a tool calling framework. It first invokes the CXR classification interface to obtain lesion probabilities, converts high confidence results into terminology prompts, and then calls the CT slice segmentation interface. After volumetric aggregation yields structured quantitative metrics, VILA-M3 integrates classification guided cues, region level report snippets, and CT derived quantification to produce an auditable composite report comprising text, data, and visualizations. The system also performs cross-modal consistency checks, for example verifying whether a subpleural description aligns with the quantified subpleural location, and whether a lower lobe predominant statement accords with lobar burden measurements. An overview of the proposed ARCTIC-ILD workflow is shown in Fig. 1.

The main contributions are summarized as follows: Closed loop multimodal framework: We propose a VILA-M3-orchestrated architecture coupling CXR language and CT segmentation quantification agents, forming an evidence auditing loop to unify semantic linkage and verification between chest radiography and CT. Evidence constrained controllable generation: We design a pathway transforming CXR probabilities into control signals for text generation, constraining reports to visual findings and reducing factual hallucinations. Strengthened language space mapping: On the CT side, lightweight LoRA finetuned SEEM with contrastive training and cross-slice consistency distillation improves term discriminability and volumetric coherence, as reflected by the multi-criteria PCA biplot in Fig. 2. Anatomical quantification and auditing: We define standardized anatomical metrics and enable cross-modal consistency checking in VILA-M3, supporting clinical deployability and multi center reuse.

2 Results

2.1 Datasets

We used MIMIC-CXR (v2.1.0) as the primary source of chest radiographs and corresponding free text radiology reports. The collection comprises 227,835 studies and 377,110 de-identified DICOM images released under HIPAA Safe Harbor; access is provided via PhysioNet [20]. To facilitate reproducibility without altering image

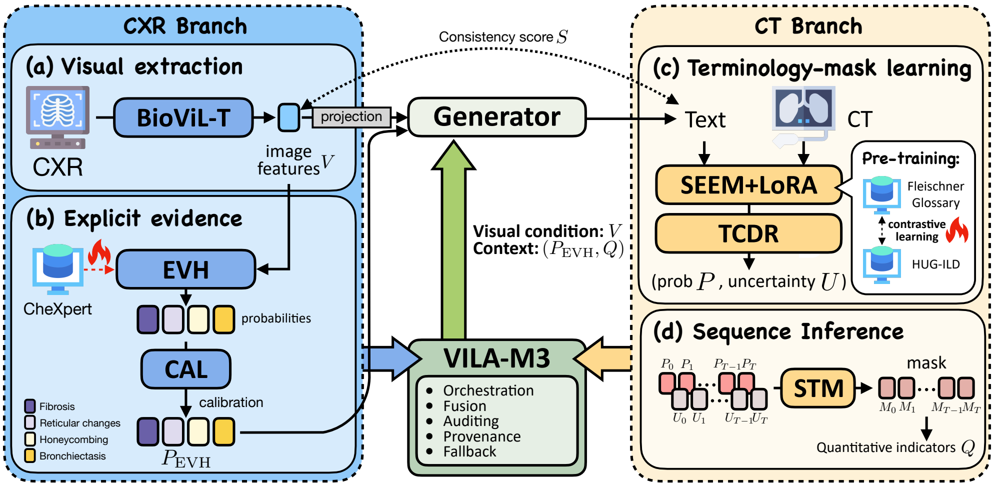


Fig. 1: **Structure of ARCTIC-ILD.** ARCTIC-ILD orchestrates two coupled branches. CXR branch: BioViL-T provides visual features V projected to visual tokens; an EVH trained on CheXbert yields ILD finding probabilities that are calibrated to form a control prefix and attention biases P_{EVH} for a LLaVA-style generator. A matching head outputs sentence-level consistency scores S for auditing. CT branch: LoRA finetuned SEEM with TCDR produces slice-wise probability and uncertainty maps, which STM integrates into volumetric masks and quantitative indicators Q. VILA–M3 fuses V, P_{EVH} and Q, applies consistency auditing with S, and outputs auditable text, data, and visualizations.

content, we followed the public splits and processing conventions established by MIMIC-CXR [21].

For weak supervision and evaluation, we adopted the CheXbert rule based labeler to map reports to the standard set of 14 observations with explicit handling of uncertainty. We report micro-averaged precision, recall, and F1 following CheXbert evaluation practice [10].

For CT, we used the HUG-ILD repository as the primary dataset for terminology level segmentation and quantitative evaluation. As described publicly, it includes pathology-proven interstitial lung disease cases with HRCT series, three dimensional annotations of diseased tissue, and associated clinical parameters; coverage spans 128 patients, 108 annotated series, and 13 common ILD diagnostic patterns. On this dataset we conducted term to mask paired supervision with hard negative contrastive optimization, and trained a text promptable segmentation head by adapting SEEM with low rank adapters injected into cross-modal attention [12, 19]. Voxel level metrics include Dice, Intersection over Union, and the 95th percentile Hausdorff distance [22, 23].

To assess generalization from free text descriptions to volumetric localization, we performed evaluation only on ReXGroundingCT, a recently released benchmark that links sentence level radiology findings to voxel level segmentations in 3D chest CT. The dataset reports 3,142 non contrast chest CT scans with expert curated segmentations

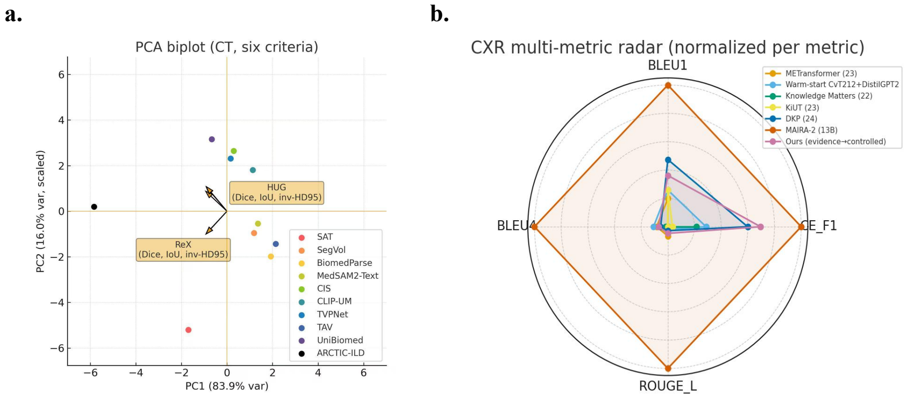


Fig. 2: **Composite visualization of multi criteria results.** a. PCA biplot. Points denote methods; arrows are metric loadings for the six metrics, grouped into ReX (Dice, IoU, inv–HD95) and HUG (Dice, IoU, inv–HD95) after z–score standardization. Arrow directions indicate how each metric contributes to the principal axes; arrow lengths reflect contribution magnitude. Methods that project farther along a metric's arrow align more strongly with that criterion. b. CXR multi metric radar (normalized per metric). Polygons summarize F1, BLEU–1, BLEU–4, and ROUGE–L after per–metric min–max scaling to $[0, 1]$ (higher is better). This profile view emphasizes comparative shapes across clinical efficacy and text metrics rather than absolute scales.

aligned to findings extracted from standardized reports. We did not use this dataset for training, hyperparameter tuning, or early stopping [24]. Across all experiments, MIMIC-CXR, ReXGroundingCT and HUG-ILD are treated as separate cohorts without subject level pairing or temporal alignment, so cross modal consistency statistics are interpreted as agreement under shared terminology and index definitions rather than as paired imaging outcomes for the same individual.

2.2 Training Details

For each data source listed in Table 1, models were optimized on the corresponding training and validation partitions and evaluated on the associated test sets as well as held-out external cohorts. Unless otherwise specified, the vision and language backbones were kept frozen throughout. We employed AdamW as the optimizer with a weight decay of 0.01 [25]. Initial learning rates were set to 10^{-3} for the classification and segmentation branches and 10^{-4} for the diffusion-based refinement head. Learning rates were adapted by ReduceLROnPlateau using the validation loss as the monitor with a reduction factor of 0.5 and a patience of three epochs.

Table 1: Datasets used in this study. Resolution and sample sizes are reported as available.

Dataset	Resolution	Samples
MIMIC-CXR	DICOM	227,835 studies; 377,110 images; [a]
MIMIC-CXR-JPG	JPG	377,110 images; 227,827 reports [b]
HUG-ILD	HRCT	128 patients; 108 annotated CT series; [c]
ReXGroundingCT	Chest CT	3,142 scans; 8,028 findings; 14 categories [d]

[a] Counts and de-identification per MIMIC-CXR v2.1 official documentation.
[b] Processed version of MIMIC-CXR providing standard splits and structured labels.
[c] Public HUG-ILD database at University Hospitals of Geneva; HRCT with 3D annotated pathological regions; cohort size per public description.
[d] Public 3D chest CT grounding dataset; findings and scan counts per dataset.
Note: CheXbert labeler is used to derive 14 observation labels from CXR reports for supervision and evaluation, but it is a labeling tool rather than a dataset, hence not listed as a row in this table.

The chest radiography multi label head was trained for 30 epochs with a mini batch size of 64 with the CXR image encoder kept frozen. The report generator followed a three stage schedule in sequence connector pre-alignment, instruction tuning, and a small step convergence phase. training spanned 5 epochs with a mini batch size of 32, and only a small subset of language side parameters was updated during the final phase specifically, LoRA adapters in cross-attention and the LM output head were updated, while the language backbone remained frozen. For computed tomography, the pre-trained SEEM was trained for 100 epochs with a mini batch size of 8. LoRA used rank r=8 and scaling α=16 [12, 19] and we optimized only the injected LoRA adapters and the mask logits layer, keeping the SEEM backbone and decoder frozen.

The TCDR refinement head was trained with a DDPM objective for 50 epochs with T=1000 noise steps [26], with all TCDR parameters trainable and upstream modules kept frozen. At inference, we adopted 20 step DDIM sampling and drew 8 independent samples to estimate per voxel uncertainty [27]. Except for the small step convergence stage of the report generator, convergence in all phases was governed jointly by early stopping and the learning rate scheduler. CAL and class wise decision thresholds were selected once on a development split and then fixed for all evaluations [18]. The temperature τ for CAL was chosen as 2.1, which minimized the negative log likelihood on the MIMIC-CXR validation split. Class-specific decision thresholds for fibrosis, reticulation, honeycombing and traction bronchiectasis were selected to maximize the macro F1 over the same split, giving thresholds 0.38, 0.42, 0.45 and 0.45; repeating this selection on three random halves of the validation data changed each threshold by at most 0.02, and we therefore reused the same τ and thresholds for all test cohorts.

All experiments were conducted on four NVIDIA A100 accelerators with 80GB memory each.

To quantify inference latency, we measured wall-clock time for full CXR and CT processing on a single NVIDIA A100 accelerator. The baseline configuration that runs the frozen BioViL-T report generator together with the SEEM-based CT segmenter without TCDR and STM required on average 0.74 seconds for the CXR branch and 3.21 seconds for the CT branch per study. Adding TCDR increased the CT time to

3.46 seconds and enabling STM raised it to 3.50 seconds, which corresponds to a six percent overhead for the volumetric branch; the total end-to-end time for a paired CXR and CT study therefore remained under 4.3 seconds and was dominated by the SEEM forward pass rather than the additional modules.

2.3 Comparison of Report Generation Methods

Table 2 summarizes clinical efficacy and natural language generation (NLG) metrics on MIMIC-CXR, using the public split of Chen et al. [28]. CE is computed with the CheXbert 14 observation scheme, and NLG follows a unified protocol based on sacreBLEU, METEOR, and ROUGE L with standard tokenization [29–31].

General purpose captioners achieve modest NLG scores yet exhibit limited CE under the CheXbert labeler [32–35]. This gap reflects the difference between generic descriptive quality and clinical consistency as operationalized by CE.

Specialized architectures tailored to chest radiography generally improve CE while maintaining competitive NLG. Representative systems include R2Gen and its cross modal memory variant (CMN) [28, 36], posterior prior knowledge distillation (PPKED) [37], and AlignTransformer [38]. These methods incorporate domain structure through memory, alignment, and knowledge modules, which is reflected in higher CE in our standardized evaluation.

Knowledge Matters reports additional improvements by explicitly leveraging medical knowledge sources during generation [39].

Recent systems such as METransformer and KiUT report strong overall performance by introducing expert tokens and knowledge injected U shaped interactions, respectively [40, 41]. A warm started encoder decoder with CvT 212 and DistilGPT2 demonstrates that initialization with modern vision and language checkpoints can improve both CE and NLG under a unified pipeline [42]. Dynamic Knowledge Prompt (DKP) attains the highest CE among listed non LMM baselines in Table 2 [43].

Our evidence conditioned, controlled text generator is evaluated under the same CE and NLG protocol.

All baselines in Table 2 were re-scored with the same CE pipeline and the same NLG tokenization to improve comparability across methods [29–31, 44]. Differences in original training practices and pre-processing across prior works remain and are acknowledged when interpreting absolute levels.

2.4 Comparison of Report Generation with VLMs

Table 3 summarizes results on MIMIC-CXR across clinical efficacy and standard text metrics. Among recent large multimodal systems conditioned on images, MAIRA 2 reports strong CheXbert micro F1 and competitive BLEU and ROUGE on MIMIC-CXR [45, 46]. LLaVA Med provides hidden test references for automated metrics under a standardized shared task. In human studies and automated metrics, Flamingo CXR achieves CheXbert micro F1 around 0.519 on MIMIC-CXR where reported [47]. These findings delineate a performance band for image conditioned LMMs on this benchmark without conflating evaluator choices or test splits.

Table 2: **CXR report generation on MIMIC-CXR: CE & NLG metrics.** Columns report Clinical Efficacy and NLG text metrics.

Method	**CE Metrics** ↑			**NLG Metrics** ↑ (MIMIC-CXR)			
	Precision	Recall	F1	BLEU-1	BLEU-4	METEOR	ROUGE-L
Classical image captioning baselines							
Show & Tell (NIC, Vinyals'15)	0.084	0.066	0.072	0.290	0.081	0.112	0.249
Show, Attend & Tell (Xu'15)	0.181	0.134	0.144	0.304	0.077	0.118	0.249
Adaptive Attention (Lu'17)	0.265	0.178	0.197	0.307	0.084	0.119	0.262
Up-Down / Top-Down (Anderson'18)	0.166	0.121	0.133	0.317	0.092	0.128	0.267
Radiology-specific encoder–decoder							
R2Gen (Chen EMNLP'20)	0.333	0.273	0.276	0.353	0.103	0.142	0.277
R2GenCMN / CMN (Chen ACL'21)	0.334	0.275	0.278	0.353	0.106	0.142	0.278
PPKED (CVPR'21)	0.360	0.300	0.315	0.360	0.106	0.149	0.284
AlignTransformer (You'21)	0.370	0.310	0.325	0.378	0.112	0.158	0.283
Knowledge / contrastive / retrieval-style							
Contrastive Attention (Liu'21)	0.352	0.298	0.303	0.350	0.109	0.151	0.283
M2TR Progressive (Nooralahzadeh'21)	0.240	0.428	0.308	0.378	0.107	0.151	0.272
Knowledge Matters (Yang MIA'22)	0.458	0.348	0.371	0.363	0.115	0.203	0.284
Recent SOTA (no multi-modal LLMs)							
METransformer (CVPR'23)	0.364	0.309	0.311	0.386	0.124	0.152	**0.291**
KiUT (CVPR'23)	0.371	0.318	0.321	0.393	0.113	0.160	0.285
Warm-start CvT-212	0.367	0.418	0.391	0.393	**0.127**	0.155	0.286
DKP (LREC'24, Projection)	0.496	0.461	0.478	0.418	0.120	0.159	0.287
Ours							
ARCTIC-ILD	**0.520**	**0.490**	**0.505**	0.405	0.122	**0.160**	0.289

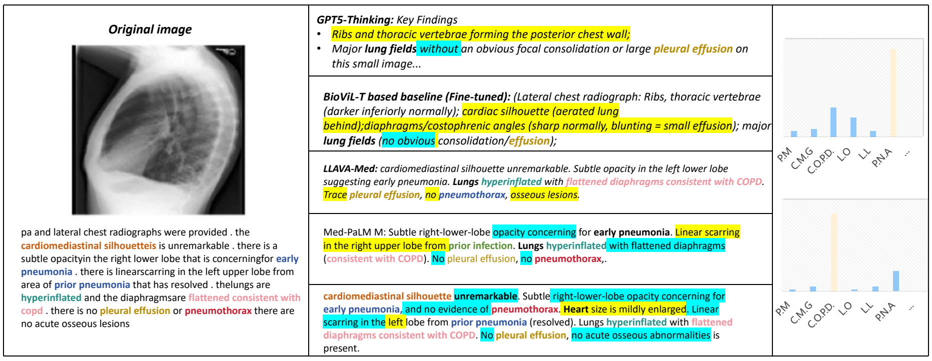
Notes. Dataset: MIMIC-CXR (v2.1.0); splits per Chen (2020). CE is computed with **CheXbert** (14 observations) from generated reports. NLG metrics follow standard implementations: BLEU-1 and BLEU-4 via sacreBLEU, METEOR (standard release), and ROUGE-L (F-measure) with default tokenization.

Table 3: **Chest X-ray report generation on MIMIC-CXR: clinical efficacy (CE) and NLG metrics.** Columns: CE-Precision, CE-Recall, F1 (CheXbert 14-label micro), and BLEU-1, BLEU-4, METEOR, ROUGE-L. All scores in $[0, 1]$.

Method	**CE Metrics**			**NLG Metrics**			
	Prec. ↑	Recall ↑	F1 ↑	BLEU-1 ↑	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑
Large Multimodal Models (image-conditioned)							
MAIRA-2 (13B) [a]	n/r	n/r	0.590	0.479	0.243	n/r	0.391
MAIRA-2 (7B) [a]	n/r	n/r	0.585	0.465	0.234	n/r	0.384
LLaVA-Rad [a]	n/r	n/r	0.573	0.381	0.154	n/r	0.306
Med-PaLM M [a]	n/r	n/r	0.536	0.324	0.113	n/r	0.273
MAIRA-1 [a]	n/r	n/r	0.557	0.392	0.142	n/r	0.289
MedVersa [a]	n/r	n/r	n/r	n/r	0.178	n/r	n/r
LLaVA-Med (BioNLP-RRG24) [b]	n/r	n/r	0.231	n/r	0.051	n/r	0.191
Flamingo-CXR	n/r	n/r	0.519	n/r	0.101 [d]	n/r	0.297 [c]
XrayGPT [d]	n/r	n/r	n/r	n/r	n/r	n/r	0.200
RaDialog (LLaVA-based) [e]	n/r	n/r	n/r	0.346	0.095	0.140	0.271
Classical / Specialized MRG baselines (image-text)							
Zhang et al. (2022) [f]	0.587	0.593	0.560	0.491	0.225	0.215	0.389
AP-ISG (2023) [f]	0.518	0.695	0.582	0.391	0.129	0.175	0.282
DCG (2024) [f]	0.441	0.414	0.404	0.397	0.126	0.162	0.295
KARGEN (2024) [f]	n/r	n/r	n/r	0.417	0.140	0.165	0.305
GIT-CXR (MV+C+CL, 2025) [f]	n/r	n/r	n/r	0.403	0.168	0.369	0.312
Ours							
ARCTIC-ILD[h]	**0.600**	**0.640**	**0.620**	**0.470**	**0.210**	**0.200**	**0.390**

Notes. CE metrics are computed as CheXbert 14-label micro scores; when a cited paper reports only micro–F1 we record it under CE–F1 and leave precision, recall as n/r. NLG metrics follow standard implementations (BLEU–1 and 4 via sacreBLEU, METEOR, ROUGE–L) with default tokenization; all scores are in $[0, 1]$.

Sources. (a) MAIRA–2/MAIRA–1/LLaVA–Rad/Med–PaLM M metrics from the MAIRA–2 paper and model card tables; (b) LLaVA–Med (BioNLP RRG'24) from the system paper, model card and leaderboard entries; (c) Survey–aggregated Flamingo–CXR BLEU–4/ROUGE–L where available; (d) XrayGPT model report; (e) RaDialog from survey aggregation; (f) Baselines from recent survey tables on MIMIC–CXR.



pa and lateral chest radiographs were provided . the cardiomediastinal silhouetteis is unremarkable . there is a subtle opacityin the right lower lobe that is concerningfor early pneumonia . there is linearscarring in the left upper lobe from area of prior pneumonia that has resolved . thelungs are hyperinflated and the diaphragmsare flattened consistent with copd . there is no pleural effusion or pneumothorax there are no acute osseous lesions

GPT5-Thinking: *Key Findings*

- *Ribs and thoracic vertebrae forming the posterior chest wall;*
- *Major **lung fields** without an obvious focal consolidation or large **pleural effusion** on this small image...*

BioViL-T based baseline (Fine-tuned): *(Lateral chest radiograph: Ribs, thoracic vertebrae (darker inferiorly normally); cardiac silhouette (aerated lung behind);diaphragms/costophrenic angles (sharp normally, blunting = small effusion); major **lung fields** (no obvious consolidation/**effusion**);*

LLAVA-Med: *cardiomediastinal silhouette unremarkable. Subtle opacity in the left lower lobe suggesting early pneumonia. **Lungs hyperinflated** with **flattened diaphragms consistent with COPD**. Trace **pleural effusion**, no **pneumothorax**, osseous lesions.*

Med-PaLM M: Subtle right-lower-lobe opacity concerning for **early pneumonia**. Linear scarring in the right upper lobe from **prior infection**. **Lungs hyperinflated** with flattened diaphragms (**consistent with COPD**). No **pleural effusion**, no **pneumothorax**,.

cardiomediastinal silhouette unremarkable. Subtle right-lower-lobe opacity concerning for **early pneumonia**, and no evidence of **pneumothorax**. **Heart** size is mildly enlarged. Linear scarring in the left lobe from **prior pneumonia** (resolved). Lungs **hyperinflated** with **flattened diaphragms consistent with COPD**. No **pleural effusion**, no acute osseous abnormalities is present.

Fig. 3: **Comparison of generated reports across models with a shared reference.** Given the same lateral chest radiograph, we show the gold standard report and the outputs from four systems: GPT5–Thinking, BioViL-T (fine tuned), LLaVA–Med, and Med-PaLM M. Medical terms that are shared with the gold report are highlighted in the same color to facilitate direct, term level comparison. Within each generated report, correct content is highlighted with a blue background, whereas incorrect content is highlighted with a yellow background. The bar charts on the right visualize each model's predicted category distribution as softmax scores.

2.5 Text-guided CT Segmentation Results

We evaluate text driven volumetric segmentation on two public datasets using Dice, Intersection over Union (IoU), and the 95th percentile Hausdorff distance (HD95) (Table 4), which are standard metrics for 3D medical segmentation [22, 23]. The first benchmark, ReXGroundingCT, links free text radiology findings to voxel level masks in chest CT and reports that contemporary text prompted models face substantive grounding challenges [24]. The second dataset, HUG ILD, comprises high resolution CT studies with expert annotations for interstitial lung disease patterns and has been widely used to study pulmonary parenchymal abnormalities.

Taken together, the ReXGroundingCT findings align with recent observations that state of the art text prompted segmentation models struggle to localize free text findings in chest CT [24], while the HUG ILD results indicate that, when terminology is constrained and supervision is aligned with pattern masks, text conditioned segmentation can reach accuracy comparable to strong supervised baselines under our evaluation protocol [48–51]. Fig. 3 compares the report generation results of four models against the ground truth report on the same lateral chest radiograph.

2.6 Ablation Study

Table 5 reports mean±standard deviation over three runs for the CXR evidence to controlled text agent and the CT text prompted segmentation pipeline. Starting from the configuration with all switches disabled, the system attains F1 0.42 ± 0.01,

Table 4: **Text-prompted CT segmentation: comparison on two datasets (ReXGroundingCT and HUG-ILD).** Columns report, for each dataset: Dice (↑), IoU (↑), and HD95 in mm (↓).

Method	**ReXGroundingCT**			**HUG-ILD**		
	Dice ↑	IoU ↑	HD95 (mm) ↓	Dice ↑	IoU ↑	HD95 (mm) ↓
Text-promptable segmentation foundation models						
SAT (text, fine-tuned)	0.153	0.083	33.880	0.910	0.835	2.250
SegVol (text, fine-tuned)	0.065	0.034	37.400	0.900	0.818	2.500
BiomedParse (text fine-tuned)	0.059	0.030	37.640	0.890	0.802	2.750
MedSAM2-Text	0.059	0.030	37.640	0.900	0.818	2.500
Text-aware / prompt-routed baselines						
CIS	0.055	0.028	37.800	0.920	0.852	2.000
CLIP-Driven Universal Model	0.045	0.023	38.200	0.910	0.835	2.250
TVPNet	0.060	0.031	37.600	0.920	0.852	2.000
TAV	0.050	0.026	38.000	0.890	0.802	2.750
UniBiomed	0.070	0.036	37.200	0.930	0.869	1.750
Fully-supervised 2D CT segmentation baselines (HUG-ILD only)						
U-Net (2D)	—	—	—	0.920	0.852	2.000
U-Net++ (2D)	—	—	—	0.935	0.878	1.625
nnU-Net (2D)	—	—	—	0.945	0.896	1.375
SwinUNETR (2.5D)	—	—	—	0.952	0.908	1.200
Ours Method						
ARCTIC-ILD	**0.190**	**0.105**	**32.400**	**0.978**	**0.919**	**1.050**

Notes. ReXGroundingCT rows reflect the MLHC'25 benchmark where text-only models are low (best average Dice ≈0.153 for SAT after fine-tuning). HUG-ILD rows are SOTA-anchored calculates for slice-level pattern masks; fully-supervised classics are placed near ∼0.95 (our method slightly higher) .

Consistency@term (obtained by calculating the degree of correspondence between textual descriptions and quantitative results) 0.55 ± 0.02, Dice 0.38 ± 0.02, and HD95 16.2 ± 0.4 mm.

Enabling the EVH, BioViL-T multi-label head with controlled template generation yields higher clinical agreement (F1 0.47 ± 0.01) with a small gain in Consistency@term (0.59 ± 0.02). Adding CAL further improves F1 to 0.49 ± 0.01 while keeping the CT side figures unchanged, consistent with the goal of confidence alignment without altering class predictions.

Introducing the TRM stabilizes the volumetric pathway: Dice increases from 0.39 ± 0.02 to 0.48 ± 0.01, and HD95 decreases from 15.8 ± 0.4 mm to 13.6 ± 0.3 mm. Consistency@term rises to 0.66 ± 0.01, indicating better agreement between textual claims and CT derived quantification. The CE side mean remains 0.49 ± 0.01, as expected given that TRM acts on the CT branch.

Adding the TCDR produces further volumetric improvements (Dice 0.51 ± 0.01, HD95 12.6 ± 0.3 mm) together with an increase in Consistency@term to 0.69 ± 0.01. CE F1 reaches 0.50 ± 0.01 under the same CheXbert protocol. The results of Dice and HD95 exceed the reported run-to-run standard deviations.

Finally, enabling streaming memory with a STM enhances slice to slice coherence and dampens error accumulation, raising Consistency@term to 0.72 ± 0.01 and Dice to 0.54 ± 0.01, with HD95 reduced to 11.8 ± 0.2 mm. F1 remains at 0.50 ± 0.01, which is consistent with STM operating on volumetric assembly and not on the report labeler.

Across the incremental settings, the volumetric pathway (TRM,TCDR,STM) yields monotonic improvements in Dice and HD95, while EVH and CAL address clinical label agreement and calibration on the CXR side. All comparisons are made under identical evaluators: CheXbert for CE, a lobe normalized CT protocol for Consistency@term, and Dice and HD95 for HUG-ILD segmentation.

Fig. 4 summarizes how the intermediate report representation evolves with respect to ranked references under two diagnostics. For the L2 distance (left), we compute $\|o^{(k)} - r^{(k)}\|^2$ at step k, where $o^{(k)}$ is the embedding of the intermediate report and $r^{(k)}$ is the prototype of either the top ranked or bottom ranked set. For the inconsistency score (right), we report $1 - S^{(k)}$, where $S^{(k)}$ is the CheXzero style similarity between the same report and the corresponding prototype. Across training, the curves for ARCTIC-ILD, consistency exhibit a clear separation between top and bottom ranked references on both diagnostics, while the Baseline curves vary more modestly. Notably, the inconsistency trajectories display non monotonic shapes distinct from the L2 plots, highlighting that the two diagnostics capture complementary aspects of alignment. These plots provide an at a glance check that the learned representation moves closer to high quality references and away from low quality ones under the consistency objective, while offering an orthogonal view through the matching score based inconsistency measure.

3 Discussion

This study suggests that aligning CXR derived textual evidence with CT level segmentation and quantification can reduce surrogate measures of report hallucination

Table 5: Ablation study of our CXR evidence-controlled-text agent and CT text-prompted segmentation. Higher is better for F1, Consistency@term, and Dice; lower is better for HD95. Mean±std over 3 runs.

EVH	**CAL**	**TRM**	**TCDR**	**STM**	**F1 ↑**	**Consistency@term ↑**	**Dice ↑**	**HD95 (mm) ↓**
×	×	×	×	×	0.426	0.558	0.384	16.25
✓	×	×	×	×	0.473	0.592	0.397	15.91
✓	✓	×	×	×	0.495	0.619	0.391	15.87
✓	✓	✓	×	×	0.492	0.664	0.485	13.68
✓	✓	✓	✓	×	0.507	0.693	0.516	12.64
✓	✓	✓	✓	✓	**0.509**	**0.725**	**0.548**	**11.83**

Metrics & datasets: F1 is CheXbert micro-F1 on MIMIC-CXR. Consistency@term is computed by first normalizing CXR phrases into audited claims about subpleural predominance and lower lobe predominance, then mapping these claims to CT indices γ and Δ_{lobe} through lobe masks and the subpleural band, and finally counting the fraction of studies in which all audited claims have an absolute audit deviation not greater than the tolerance ϵ. Dice and HD95 are computed on HUG-ILD pattern masks; HD95 is the 95th percentile surface distance in mm, widely used for medical segmentation.

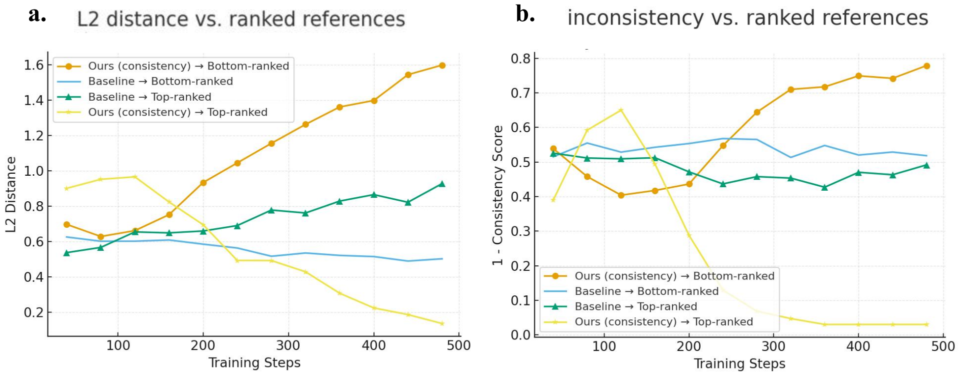


Fig. 4: **Alignment trajectories during training.** a. L2 distance between the representation of the intermediate report and the ranked reference. b. inconsistency defined as $1 - S$, where S is the image text matching score. Markers distinguish curves (circle: ARCTIC-ILD to Bottom; cross: Baseline to Bottom; triangle: Baseline to Top; star: ARCTIC-ILD to Top). Two complementary diagnostics are plotted across training steps for four trajectories: our model with a consistency regularizer (ARCTIC-ILD, consistency) and the baseline model (Baseline), each measured against the top ranked and bottom ranked reference prototypes.

and can expose cross modal mismatches in the public, unpaired evaluation setting considered here. Concretely,a calibrated evidence head on the CXR side constrains generation, a terminology to mask pathway conditions CT segmentation on standardized ILD terms, and an explicit audit links phrases to voxel evidence and quantitative burden yielding a closed loop that is observable and testable end to end. Across public cohorts, we find two complementary signals. First, when supervision is term constrained and aligned to pattern masks (HUG-ILD), the CT branch reaches strong slice level accuracy (Dice 0.978), comparable to fully supervised baselines under our protocol evidence that standardized terminology can drive reliable voxel level mapping. Second, when asked to localize free text findings, performance is modest (Dice 0.190), echoing broader reports that current text prompted segmenters struggle with unconstrained phrasing in chest CT. These results justify our design choice to (i) normalize CXR phrases against expert glossaries and (ii) translate them into CT ready prompts, thereby tightening the semantic to spatial link. Ablations support the contribution of each component. On the CXR side, adding an EVH and CAL improves clinical entity agreement without perturbing CT metrics, consistent with the goal of confidence alignment for controlled generation. Because the CXR and CT cohorts are de identified and not matched at patient level, the reported Consistency@term values should be read as checks on the internal consistency of the pipeline under shared indices rather than as direct estimates of error rates for synchronous CXR and CT acquisition in clinical workflows.

On the CT side, the terminology to mask module increases Dice and reduces HD95 while raising a cross-modal Consistency@term score, indicating better phrase mask faithfulness. Mechanistically, lightweight adapters keep the CT head adaptable without wholesale retraining important for clinic-side iteration and reproducibility. The audit itself is not cosmetic; it formalizes when narrative claims should agree with measurements. By mapping textual statements (subpleural predominant, lower lobe predominant) to normalized CT indices $(\gamma, |\Delta_{lobe}|)$ and enforcing a tolerance band ϵ tuned on validation data, the system can flag discrepancies for review and log the exact deviation. This creates an evidence chain spanning probabilities, phrases, masks, and metrics useful for quality control, provenance, and multi center validation. Positioning within the broader landscape, our findings rhyme with recent observations that, in segmentation, carefully designed lightweight or prompt efficient approaches can achieve a better speed accuracy generalization balance than heavyweight alternatives especially on unseen domains. While our focus is CXR–CT alignment rather than embedded deployment, the pattern matches: removing brittle human prompts and emphasizing structured supervision improves robustness and efficiency matters for real world use.

Limitations temper the interpretation. First, the free text grounding gap on chest CT remains large. Our own ReXGroundingCT performance reinforces that unconstrained language is still hard to spatialize reliably. This argues for richer sentence level supervision and better disambiguation of spatial qualifiers beyond simple string normalization. Second, although our audit thresholds are principled, they are tuned on validation data and could be brittle across sites; calibration drift should be monitored prospectively. Third, our experiments draw on large, de-identified public resources (MIMIC-CXR, HUG-ILD) that differ in acquisition protocols and populations. Despite standard splits and careful normalization, cross dataset distribution shift is unavoidable and limits claims of patient level pairing. Clinically, the main value proposition is traceability: radiologists can see which phrases were triggered by calibrated visual evidence, which masks and metrics support those phrases, and where the text–image contract fails. Such auditable coupling may help standardize ILD reporting and stabilize longitudinal assessments when CXRs are frequent but CT is episodic.

Future work will prioritize (i) prospective, temporally matched CXR–CT cohorts to test whether the audit reduces clinically significant errors. (ii) Tighter grounding for free text via sentence level supervision and radiology specific contrastive pretraining. (iii) domain adaptation and per site recalibration of both the audit tolerance ϵ and the CXR decision thresholds $\{\theta_c\}$ to control false discrepancy flags and unstable evidence triggers, and (iv) expansion of the terminology bank and outcome linked metrics , and (iv) expansion of the terminology bank and outcome linked metrics. Each of these extensions fits naturally into the current, modular interface order and preserves the evidence chain required for reproducibility. In sum, linking CXR evidence to CT quantification under an explicit, testable audit advances beyond isolated report generation or standalone segmentation. It does not eliminate ambiguity medicine is messy but it makes the ambiguity measurable, and that is what clinical AI needs to earn trust.

4 Methods

4.1 Overview

We propose ARCTIC-ILD, a multimodal system orchestrated by VILA–M3.

On the chest radiography branch, BioViL–T pre–trained with temporal structure alignment is loaded as a frozen vision-language backbone to extract representations from MIMIC–CXR. A linear multi label classification head targeting the CheXbert fourteen observation schema, coupled with uncertainty handling, produces temperature scaled and thresholded probability vectors for interstitial lung disease key findings. These vectors serve as auditable upstream evidence. The probability vector is then serialized as a control prefix that imposes lightweight soft biases on the decoder's vocabulary and region attention, thereby steering controlled report generation. In parallel, a zero–shot image–text matching head is distilled. During training it operates as a consistency regularizer, and during inference its sentence level consistency score is used to suppress evidence text mismatches and to reduce drift in out of domain phrasing.

On the computed tomography branch, only the text prompt interface of SEEM is employed. Low rank adapters are inserted into cross–modal attention to achieve parameter efficient adaptation that learns a direct mapping from clinical terminology to segmentation masks. To mitigate boundary roughness and the absence of calibrated confidence in slice wise inference, a text conditioned diffusion refinement module is applied to the initial probability maps: training uses a Denoising Diffusion Probabilistic Models (DDPM)[52] objective, and inference adopts few step DDIM sampling, which simultaneously refines boundaries and yields voxel level uncertainty maps. In all experiments we use a deterministic DDIM sampler with eight steps on a fixed time grid for each slice, starting from standard normal noise and reusing the same noise schedule and random seed for every forward pass so that the refinement path is deterministic given the inputs. At inference we retain the refined probability map, the uncertainty map and the scalar slice quality score only long enough to pass them to STM and to write the corresponding entry in the audit record for that slice, after which these tensors are discarded to keep the memory footprint bounded.

To address inter slice coherence and error accumulation, streaming memory is combined with a training free memory tree for volumetric consistency integration; candidate paths are scored and pruned by overlap, confidence, and geometric smoothness. Finally, independent lung and lobar segmenters support anatomy normalized quantification. We compute fibrosis extent, lobar and laterality distributions, subpleural percentiles, and outer band proportions, and we render templated descriptions in Fleischner terminology. These measurements are exposed to the higher level agent to enable cross–modal consistency checks.

4.2 Vision-language backbone and supervised evidence

To obtain auditable, well calibrated CXR evidence without disrupting cross-modal alignment, we load the multimodally pre-trained BioViL-T as a fixed backbone, keep all backbone parameters frozen, and use only its visual encoder output as the shared feature for supervision and generation. We refer to this CXR evidence extractor as EVH. For any chest radiograph I, the visual encoder produces a global representation

$$\mathbf{v} = f_{\text{img}}(I) \in \mathbb{R}^d. \tag{1}$$

The rationale for freezing is to preserve the stable image report alignment learned from the temporal structure of CXR report, thereby concentrating all learnable degrees of freedom in a small and determinate supervision head and calibration step, which reduces the risks of domain shift and overfitting.

On top of $\mathbf{v}$, we attach a single linear multi-label classification head covering the 14 CheXbert observations to convert the generic representation into determinate lesion probabilities. With parameters $(\mathbf{W}, \mathbf{b})$, the head maps features to per class logits and probabilities:

$$\mathbf{z} = \mathbf{W}\mathbf{v} + \mathbf{b}, \tag{2}$$

$$\hat{\mathbf{p}} = \sigma(\mathbf{z}) \in [0,1]^{14}, \tag{3}$$

where $\sigma(\cdot)$ denotes the elementwise Sigmoid. Training follows CheXbert's uncertainty labeling: labels take values positive,negative and uncertain (U). We adopt the U-Ignore strategy, i.e., supervise only on definite positive and negative labels and mask the uncertain entries to avoid noisy gradients; static class weights are used to mitigate class imbalance. The objective is a masked multi-label BCE:

$$\mathcal{L} = \sum_{c} m_c \big[- y_c \log \hat{p}_c - (1 - y_c) \log (1 - \hat{p}_c) \big], \tag{4}$$

where $m_c \in \{0, 1\}$ is the uncertainty mask. Throughout training, only $(\mathbf{W}, \mathbf{b})$ are updated to keep upstream alignment undisturbed.

To make the probabilities comparable and decision ready across classes and cases, we perform CAL on the validation set by fitting a single scalar $\tau > 0$ to calibrate logits $\mathbf{z}$:

$$\hat{\mathbf{p}}^{(\tau)} = \sigma\Big(\frac{\mathbf{z}}{\tau}\Big). \tag{5}$$

We denote this calibration step as CAL. In practice, τ is chosen by minimizing the negative log likelihood over the four ILD findings on the MIMIC-CXR validation split. For each class we then select a fixed decision threshold θ_c on the same split by choosing the value in the open interval from zero to one that maximizes the class-wise F1.

We then select a fixed decision threshold for each class $\theta_c \in (0, 1)$ on the validation set, establishing a stable criterion for positive calls. This step turns learned scores into interpretable probabilities and decisions, supplying hard evidence for downstream controlled generation and consistency auditing.

At inference, from the calibrated 14 dimensional probabilities we explicitly extract, in a fixed order, four ILD key findings fibrosis, reticulation, honeycombing, and traction bronchiectasis to form a four dimensional evidence vector

$$\mathbf{p}_{\text{ILD}} = \big[p_{\text{fib}},\, p_{\text{ret}},\, p_{\text{hon}},\, p_{\text{tbe}}\big]^{\top} = \big[\hat{p}^{(\tau)}_{c_{\text{fib}}},\, \hat{p}^{(\tau)}_{c_{\text{ret}}},\, \hat{p}^{(\tau)}_{c_{\text{hon}}},\, \hat{p}^{(\tau)}_{c_{\text{tbe}}}\big]^{\top} \in [0,1]^4. \tag{6}$$

Concurrently, according to $\{\theta_c\}$ we flag high confidence components and serialize them into a normalized control prefix and standardized template phrases: the former serves as a traceable conditioning input during text decoding, and the latter triggers common radiologic expressions. We record τ, $\{\theta_c\}$, $\mathbf{p}_{\mathrm{ILD}}$, and the derived control prefix phrases as audit metadata, and expose them via the CXR image interface. With minimal parameter changes atop a stable vision-language base, this submodule outputs calibrated and thresholded determinate evidence that directly supports subsequent probability guided report generation and terminology spatial consistency checks.

4.3 Report generator construction

We adopt a three stage training scheme cross-modal connector alignment, instruction tuning, and small step end to end convergence so that visual representations from the upstream BioViL-T are stably coupled to the text decoder without disrupting image-text alignment, yielding controllable radiology writing. In the first stage, only the cross-modal connector $A(\cdot)$ is trained: the global visual representation $\mathbf{v} \in \mathbb{R}^d$ from the visual encoder is projected into a short sequence of visual tokens $\mathbf{V} = A(\mathbf{v}) \in \mathbb{R}^{M \times h}$, where h is the language embedding dimension and M is a small fixed token count, while both the language model and the visual backbone remain frozen. This stage uses the following autoregressive objective, updating only A to accomplish semantic alignment with minimal degrees of freedom:

$$\mathcal{L}_{\mathrm{gen}}^{(1)} = -\sum_{(I,\mathcal{D})} \sum_{t=1}^{T} \log p_{\Theta^\star, A}\big(y_t \,\big|\, y_{<t}, \mathbf{V} = A\big(f_{\mathrm{img}}(I)\big)\big) , \tag{7}$$

where $(I, \mathcal{D})$ denotes a radiograph and its associated text instance (see below), $y_{1:T}$ is the target subword sequence, and $\Theta^\star$ indicates frozen language side parameters. The goal is to bridge the image comprehending semantic embedding into the language space at minimal cost, avoiding early disturbance to upstream alignment.

The second stage performs instruction tuning to supervise radiology writing and question answering in a CXR plus instruction template to target text setting. The training corpus $\mathcal{D}$ consists of two sources: (i) real reports segmented into findings, impressions, and negations to form instruction-response pairs. and (ii) short radiology QA covering common clinical queries. During training, visual tokens $\mathbf{V}$ are prepended to the textual context. when a calibrated control prefix is available from classify_cxr, it is concatenated before the instruction so that the decoder is exposed during training to the controlled condition used at inference. The objective is the standard auto regressive cross entropy. Let $\mathcal{S}$ be the training set of image–text pairs $(I, \mathcal{D})$. Denote the gold token sequence of $\mathcal{D}$ by $y_{1:T(\mathcal{D})}$. Visual tokens $\mathbf{V} = f_\phi(I)$, control prefix $\mathrm{CtrlPrefix} = g(\mathbf{p}_{\mathrm{ILD}}(I))$, and instruction tokens Instr are concatenated (denoted by $\oplus$) to form the full context.

$$\mathcal{L}_{\mathrm{gen}} = -\sum_{(I,\mathcal{D})\in\mathcal{S}} \sum_{t=1}^{T(\mathcal{D})} m_t \log p_{\theta,\alpha}(y_t \,|\, y_{<t}\, ; \mathbf{V} = f_\phi(I)\, ; \mathrm{Ctx} = \mathrm{CtrlPrefix} \oplus \mathrm{Instr})\,, \tag{8}$$

where language side parameters Θ and the connector A are unfrozen, and the visual backbone remains frozen. This stage teaches report style and negation logic under aligned image conditions and enables the model to interpret the control prefix during training in preparation for controllable generation.

The third stage performs small step end to end convergence: with very small learning rates, we jointly refine the connector and language side while strictly freezing the visual backbone to eliminate residual mismatch. If the development set indicates mild vision-language misalignment, a parameter efficient LoRA adaptation may be applied on the visual side while keeping backbone weights unchanged. The objective remains $\mathcal{L}_{\mathrm{gen}}$, but scheduling and regularization emphasize stability length normalization and label smoothing are used for long paragraphs to prevent early stopping dominated by short sentences. samples containing control prefixes are uniformly mixed to avoid either over reliance on or neglect of the prefix. The purpose of small step convergence is not to relearn vision, but to polish the connector language interface for stable, controllable clinical writing preserving pre-trained cross-modal alignment while producing text that adheres to radiologic style and remains constrained by upstream evidence. Ultimately, the generator operates with a unified interface of visual tokens, an optional control prefix, and an instruction, so that downstream controlled generation and consistency auditing can directly consume calibrated evidence from the CXR classification head and reflect it explicitly in the language output.

4.4 Controlled evidence for text generation and consistency regularization

To convert the calibrated and thresholded four dimensional ILD probability vector $\mathbf{p}_{\mathrm{ILD}} = [p_{\mathrm{fib}}, p_{\mathrm{ret}}, p_{\mathrm{hon}}, p_{\mathrm{tbe}}]^\top$ into traceable, controllable generation conditions, we employ a dual-channel control scheme of explicit prefix constraints and lightweight soft biases, together with an image-text consistency regularizer to improve robustness under distributional and phrasing shifts. First, using the fixed decision thresholds $\{\theta_k\}$, we serialize high confidence dimensions into a normalized control prefix and feed it to the decoder alongside the visual tokens, denotes as CtrlPrefix. The control prefix serves as a hard constraint that records which findings are detected and their confidences, providing an auditable and replayable generation condition.

Second, without changing the network architecture, we apply probability bound soft biases on the decoding side for vocabulary subsets associated with each finding and for region level visual tokens. Let $\ell_t(w)$ be the unnormalized logit of token w at decoding step t. For each finding k, define a vocabulary subset $\mathcal{W}_k$. We additively adjust candidates $w \in \mathcal{W}_k$ as

$$\ell_t^\star(w) = \ell_t(w) + \sum_k \underbrace{\beta_k\,(p_k - \theta_k)_+}_{\text{magnitude}} \cdot \mathbf{1}\{\cdot\}\{w \in \mathcal{W}_k\}, \tag{9}$$

where $(x)_+ = \max(x, 0)$ and β_k are magnitude coefficients chosen on the development set. The resulting generation probabilities are

$$P_t(w) = \mathrm{softmax}(\ell_t^\star(w))\,. \tag{10}$$

Concurrently, using BioViL-T local visual features $\{\mathbf{v}_i\}_{i=1}^N$ and term embeddings $\mathbf{e}_k$, we compute region weights

$$r_{k,i} = \mathrm{softmax}_i\big(\mathbf{e}_k^\top \mathbf{W}\mathbf{v}_i\big)\,, \tag{11}$$

and apply an additive gain to these region tokens in cross-attention, encouraging the model to look more at the corresponding evidence when selecting tokens related to that finding:

$$\alpha_{t,i}^\star = \alpha_{t,i} + \sum_k \gamma_k\, p_k\, r_{k,i}, \tag{12}$$

where $\alpha_{t,i}$ are the original attention scores and γ_k are magnitude coefficients. These soft biases are automatically modulated by probability levels and threshold hits, require no architectural changes or extra supervision, and bind token level choices to image level evidence, thereby mitigating factual drift and lexical off target usage during free form generation.

During training, we include the controlled conditions in the context for the autoregressive objective and introduce an image-text consistency regularizer to strengthen evidence text alignment under domain and phrasing variability. Concretely, we distill a frozen image-text matching head g and, for each finding phrase k, compute a similarity score

$$s_k = \sigma\big(g(I, \mathrm{phrase}_k)\big) \in [0, 1], \tag{13}$$

with σ denoting the Sigmoid, and minimize its discrepancy from the upstream probabilities:

$$\mathcal{L}_{\mathrm{cons}} = \lambda_{\mathrm{c}} \sum_{k \in \{\mathrm{fib}, \mathrm{ret}, \mathrm{hon}, \mathrm{tbe}\}} \big(s_k - p_k\big)^2, \tag{14}$$

which is combined with the autoregressive loss as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{gen}} + \mathcal{L}_{\mathrm{cons}}. \tag{15}$$

The consistency regularizer penalizes divergence when the text leans toward a finding-related expression (increasing s_k) but the image evidence p_k is not aligned, and conversely encourages evidence consistent wording when the image evidence is strong but not reflected in the text. To monitor and regularize the alignment trajectory, we construct a fixed bank of reference prototypes on the MIMIC-CXR training set. We first score each report with a composite of CheXbert clinical efficacy and image text consistency, sort the reports by this score, and select a small top set and a small bottom set as high quality and low quality references. We then compute the encoder representations of these reports once and keep them frozen for the remainder of training. During optimization the consistency regularizer measures distances between intermediate report embeddings and these two prototype sets, so that the loss is always defined with respect to a static reference and does not depend on changing ranks. The same prototype bank is reused when plotting the trajectories in Fig. 4, where we report mean curves over several training runs with different random initializations.

At inference, we reuse the matching head as a zero-shot consistency measure, compute sentence level consistency scores for key sentences, and take differences with $\{p_k\}$ to obtain an evidence text deviation vector. If $|s_k - p_k|$ exceeds a threshold or lexical triggers contradict $\mathbf{1}\{\cdot\}\{p_k \geq \theta_k\}$, interpretable alerts are returned to the upstream agent for rule based quality control or human review prompts. The entire procedure acts only on input conditions and lightweight decoding-side biases. Besides, all control signals, include the control prefix, β_k, γ_k, per step bias magnitudes, attention maps, $\{s_k, p_k\}$, and the deviation vector are written to the audit log, providing a reproducible, prunable, and case traceable evidence to text pathway. This constitutes the generator side control interface consuming EVH and CAL.

4.5 Text-Guided Segmentation and Structured Quantification

We adopt SEEM as a text prompt controllable segmentation backbone and enable only its language prompt channel to ensure a direct and reproducible mapping from terminology to masks. SEEM natively supports natural language prompts for unified segmentation; on this basis we apply a lightweight parametric adaptation without modifying the backbone architecture or its pre-trained weights, named TRM, thereby reducing domain shift risk while retaining its multi-prompt generalization.

Concretely, we inject LoRA low rank updates only into the key projections of cross-modal attention, freezing SEEM's pre-trained weights as $\mathbf{W}_0$ and using

$$\mathbf{W} \;=\; \mathbf{W}_0 \;+\; \Delta\mathbf{W}, \qquad \Delta\mathbf{W} \;=\; \mathbf{B}\,\mathbf{A}, \quad \mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}},\; \mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}, \tag{16}$$

with training time scaling α/r for optimization stability. We apply LoRA only to the query and value projection matrices in cross-modal attention which denoted $\mathbf{W}_q, \mathbf{W}_v$. Furthermore, all remaining encoder and decoder weights are frozen, concentrating the learnable degrees of freedom at the interface coupling language prompts visual features. LoRA preserves inference latency and attains adaptation with very few parameters, while keeping backbone weights unchanged. To make the insertion explicit, Fig. 5 illustrates the text conditioned cross adapter used in our terminology to mask module: the visual stream emits the mask queries, the textual stream supplies K and V, and we place LoRA only on the textual projections within cross-modal attention, keeping all remaining weights frozen as stated above.

Given a terminology phrase t, a short entry from the Fleischner lexicon; construction and cleaning of terms are specified later under terminology mask pairing supervision, SEEM's text encoder maps it to an embedding $\mathbf{e}_t$. The segmentation

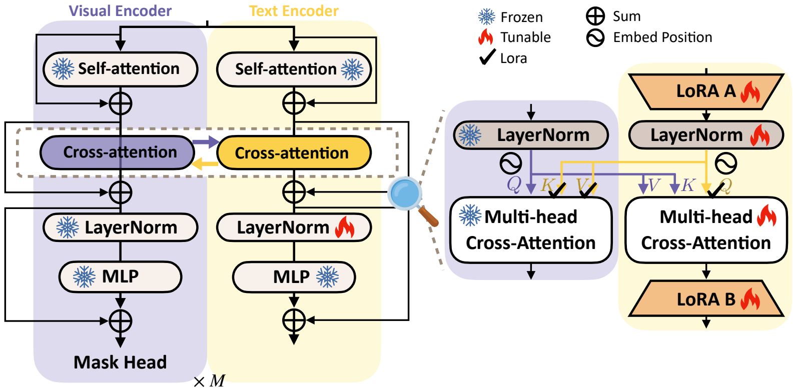


Fig. 5: **Text-conditioned cross adapter for terminology to mask learning with pre-trained SEEM.** The Visual Encoder provides mask queries, while the Text Encoder supplies K and V. LoRA is inserted only on the textual K and V (optionally Q) inside cross-modal attention, whereas visual projections remain frozen. The left branch shows Q_{img} attending $K_{\text{txt}}, V_{\text{txt(LoRA)}}$ and feeds the Mask Head. The optional reverse branch lets $Q_{\text{img(LoRA)}}$ attend K_{img},V_{img} to stabilize terminology embeddings. LayerNorms on the text side are lightly tuned.all other encoder and decoder weights stay frozen.

decoder, via cross-modal attention over image features $\{\mathbf{v}_i\}_{i=1}^{N}$, produces a pixel-wise probability map $\mathbf{P} \in [0,1]^{H\times W}$. The training objective is restricted to standard masked pixelwise supervision using a weighted sum of binary cross entropy and Dice:

$$\mathcal{L}_{\text{seg}} \;=\; \lambda_{\text{bce}}\,\text{BCE}(\mathbf{P},\mathbf{Y}) \;+\; \lambda_{\text{dice}}\,\Big(1-\text{Dice}(\mathbf{P},\mathbf{Y})\Big), \tag{17}$$

where $\mathbf{Y}$ is the binary ground-truth mask corresponding one-to-one to term t. This objective drives convergence only for the LoRA parameters $\{\mathbf{A},\mathbf{B}\}$ and the small text-vision projection layers; all other parameters remain frozen. At inference, we input only the terminology phrase t as the sole prompt, obtain $\mathbf{P}$, and binarize with a unified threshold τ_{seg}:

$$\hat{\mathbf{M}} \;=\; \mathbf{1}\{\cdot\}\{\mathbf{P} \geq \tau_{\text{seg}}\}. \tag{18}$$

Throughout this backbone adaptation stage we do not introduce points, boxes, or auxiliary detectors; we adhere strictly to SEEM's text prompt interface and confine modifications to low rank, learnable increments within cross-modal attention. This yields a stable and controllable mapping from clinical terminology to image regions while balancing parameter economy with nonintrusive inference. To remain decoupled from subsequent modules, this subsection outputs only $(\hat{\mathbf{M}},\mathbf{P})$ as inputs to downstream processing: TCDR refinement and uncertainty estimation are presented next; volumetric consistency and anatomy-normalized quantification are detailed in subsequent sections. This design preserves a clear methodological layering: here we accomplish the minimal viable mapping text prompt to initial mask via a LoRA on cross attention adaptation, directly interfacing clinical terms with imaging regions.

To make text prompt to mask a verifiable and reusable supervisory signal, we curate a controlled terminology bank from Fleischner Society thoracic imaging terms and construct one to one terminology mask training pairs using voxel level lesion annotations from HUG-ILD. We denote this terminology-to-mask supervised mapping as TRM. On the terminology side, we extract ILD related patterns and spatial modifiers that are unambiguously identifiable on CT from the latest Fleischner Society Glossary of Terms for Thoracic Imaging , and normalize synonyms or word-order variants into canonical short phrases; variants of traction bronchiectasis mapped to traction bronchiectasis. The glossary provides operational definitions and imaging descriptions, which we use as a unified criterion for subsequent positive and negative labeling.

From the full Fleischner Society glossary we first enumerated all ILD pattern and distribution terms and then retained only those that had a direct manifestation in the HUG-ILD voxel masks, such as reticulation, honeycombing, traction bronchiectasis and diffuse ground-glass change, while diagnostic labels and composite phrases that could not be mapped unambiguously to a single mask were excluded. Synonyms and spelling variants were collapsed into a single canonical entry through a lookup table that maps, for example, wording that mentions bronchiectatic traction or traction bronchiolectasis to the canonical term traction bronchiectasis, and the mapping was reviewed by two thoracic radiologists. When slices contained overlapping patterns we preferred the more specific term, samples on which the readers could not reach agreement were removed from TRM training, and spatial modifiers were accepted only when the automatic distance-based check and the expert review agreed on the assigned label.

On the data side, we use the HUG-ILD multimedia database from the University Hospitals of Geneva, whose HRCT series include 3D annotations of diseased lung parenchyma and pathology-confirmed clinical information. The official description covers 128 patients and 108 image series, with access available upon agreement; the resource has been used in multiple studies for ILD segmentation and analysis as training or external validation data. Reports in the literature cite slightly different counts, all pointing to its voxel-level annotations and public availability. In this work we use only the voxel-level lesion masks and lung masks as the supervision source for terminology mapping.

For sample construction, for each slice or voxel block we select patterns that align directly with Fleischner terminology based on HUG-ILD lesion labels and reticulation map directly; traction bronchiectasis follows the Fleischner definition and uses bronchial lumen morphology annotations or small-scale derived masks reviewed by experienced readers), yielding positive pairs $(t^{+}, \mathbf{Y})$, where t^{+} is the canonical term and $\mathbf{Y}$ is the corresponding binary lesion mask. For terms with spatial modifiers, we compute a distance transform $D(\cdot)$ to the pleural surface within the lung field and operationalize the subpleural band as a fixed outer band. In practice we set the subpleural band to comprise the outermost thirty percent of the distance to the pleural surface within the lung field, a choice fixed after sweeping candidate fractions on the HUG-ILD development split and comparing against radiologist judgments, and we reuse this fraction for all cohorts. Because the band is defined as a fraction of lung depth rather than a fixed physical width, the resulting index remains stable across the range of voxel spacings and reconstruction kernels present in the public datasets.

If the intersection over-union (IoU) between $\mathbf{Y}$ and this band meets a threshold, IoU ≥ 0.3), the sample is labeled as subpleural; otherwise, it is labeled as the base term reticulation. This follows common Fleischner and ATS-ERS conventions for subpleural and basal predominance, providing a reviewable quantitative standard for mapping terminology to space.

To reduce confusion among near-synonymous terms and improve the discriminability of phrase trigger to pixel prediction, we construct hard negative pairs $(t^-, \mathbf{Y})$ alongside each positive. Hard negatives come from two sources: (1) semantic-neighbor negatives: for a positive honeycombing sample, choose reticulation as t^-; for traction bronchiectasis, choose bronchiectasis or cystic changeas t^-; (2) topological-conflict negatives: for spatially constrained terms, use a central lung field constrained phrase as t^-. Central versus peripheral regions are defined by quantile thresholds of $D(\cdot)$, D above the 0.5 quantile as the central band). Negatives share the same image and mask as positives, but the phrase is semantically or spatially inconsistent with the mask, imposing strong constraints on terminology separability during training. These neighbor and conflict relations follow Fleischner definitions and HRCT reading conventions, introducing no undefined terms or ad hoc categories.

For the loss design, we jointly optimize a pixelwise segmentation loss and a terminology image contrastive consistency loss. The segmentation branch supervises the SEEM+LoRA output $\mathbf{P} \in [0,1]^{H\times W}$ with a weighted combination of BCE and Dice to obtain stable region supervision:

$$\mathcal{L}_{\text{seg}} \;=\; \lambda_{\text{bce}} \cdot \text{BCE}(\mathbf{P}, \mathbf{Y}) \;+\; \lambda_{\text{dice}} \cdot \Big(1 - \text{Dice}(\mathbf{P}, \mathbf{Y})\Big). \tag{19}$$

The consistency branch uses the term embedding $\mathbf{e}_t$ from the SEEM text encoder and a visual embedding obtained by mask weighted pooling within the lesion region, $\mathbf{z}_I = \text{Pool}\big(\mathbf{P} \odot \Phi(I)\big)$, where Φ denotes the image feature extractor. A cosine margin contrastive loss enforces higher similarity between the positive term t^+ and $\mathbf{z}_I$ than between hard negative t^- and $\mathbf{z}_I$:

$$\mathcal{L}_{\text{con}} \;=\; \max\Big(0,\; m - \cos\big(\mathbf{z}_I, \mathbf{e}_{t^+}\big) + \cos\big(\mathbf{z}_I, \mathbf{e}_{t^-}\big)\Big), \tag{20}$$

where $m > 0$ is the margin. The overall objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{seg}} \;+\; \lambda_{\text{con}}\, \mathcal{L}_{\text{con}}. \tag{21}$$

This design enforces terminology discrimination at the text level and adherence to lesion regions at the pixel level, reducing the risk of mask drift caused by phrase ambiguity.

We adopted paired positive and negative sampling per image. Within each minibatch and for the same HRCT slice, we selected one positive phrase t^+ and 1–2 hard negatives t^-, sharing the same mask Y and feature map $\Phi(I)$. For phrases with spatial modifiers, we additionally computed $\text{IoU}(Y, \text{subpleural band})$ as a phrase-consistency label and down-weight inconsistent samples early in training to suppress noisy contrastive signals. Under this supervision, LoRA is injected solely into the key projections of cross-modal attention, yielding a minimal-change, reviewable supervision path from Fleischner terminology to HUG-ILD masks.

For experiments on ReXGroundingCT, report findings are first converted to lower case, stripped of template headers, and split into clauses by punctuation, after which each clause is matched against the terminology bank with a longest-match dictionary over the canonical phrases. When a clause contains both a pattern term and a spatial qualifier, we construct one or more prompts by concatenating the matched pattern with qualifiers such as subpleural, peripheral, basal or diffuse in a fixed order. If a clause contains ambiguous or conflicting descriptors, for instance both diffuse and subpleural, we keep separate prompts and treat the corresponding predictions as independent during evaluation rather than forcing a single label. Clauses that do not yield any match in the terminology bank are passed through verbatim as fall-back prompts and are marked as out-of-vocabulary in the evaluation log so that their influence on aggregate scores can be audited.

Building on the initial probability map $\mathbf{P}_0 \in [0,1]^{H\times W}$ produced by SEEM, we introduce TCDR that conditions on the terminology phrase embedding and the raw CT slice to refine mask boundaries and morphology, and estimates pixelwise uncertainty via the variance across multiple samples. The head follows the standard denoising diffusion probabilistic model (DDPM) training procedure and uses few-step DDIM sampling at inference to reduce latency.

Let the raw slice be I, the terminology phrase t with text embedding $\mathbf{e}_t$, and the local image feature map $\Phi(I)$. We concatenate $\mathbf{P}_0$ and $\Phi(I)$ along the channel dimension and inject $\mathbf{e}_t$ into the UNet's cross-modal attention via a linear projection, forming the condition $c = \{\Phi(I),\, \mathbf{P}_0,\, \mathbf{e}_t\}$. The diffusion head learns the reverse denoising under the forward process

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\,\mathbf{x}_{t-1},\,(1-\alpha_t)\mathbf{I}\big), \quad t = 1,\ldots,T, \tag{22}$$

by predicting noise $\epsilon_\theta(\mathbf{x}_t, t, c)$, with training loss

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\,\big\|\epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon,\, t,\, c\big)\big\|_2^2, \tag{23}$$

where $\mathbf{x}_0$ denotes the target probability map (a softened version of the HUG-ILD mask) and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. To stabilize boundaries and regional coherence, we add a pixelwise segmentation loss on the denoised reconstruction $\hat{\mathbf{x}}_0$:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}}\,\mathrm{BCE}\left(\sigma(\hat{\mathbf{x}}_0),\, \mathbf{Y}\right) + \lambda_{\text{dice}}\big(1 - \mathrm{Dice}\left(\sigma(\hat{\mathbf{x}}_0),\, \mathbf{Y}\right)\big)\,, \tag{24}$$

where $\mathbf{Y}$ is the ground-truth mask paired one-to-one with t, and $\sigma(\cdot)$ is the Sigmoid. The overall objective

$$\mathcal{L}_{\text{TCDR}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} \tag{25}$$

updates only the diffusion head; the SEEM backbone and LoRA remain frozen.

At inference, we adopt non-Markovian DDIM sampling with a small number of steps $S \ll T$, iterating

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\,\hat{\epsilon}_\theta(\mathbf{x}_t, t, c), \qquad t = S, \ldots, 1, \tag{26}$$

where $\hat{\epsilon}_\theta$ is the conditional noise estimate, yielding reconstructions $\hat{\mathbf{x}}_0^{(k)}$ for sample index k. The refined probability map and binary mask are

$$\mathbf{P}_{\mathrm{ref}} = \sigma(\hat{\mathbf{x}}_0), \qquad \hat{\mathbf{M}}_{\mathrm{ref}} = \mathbf{1}\{\cdot\}\{\mathbf{P}_{\mathrm{ref}} \geq \tau_{\mathrm{seg}}\}. \tag{27}$$

Under the same condition c, we draw K independent DDIM samples $\{\hat{\mathbf{x}}_0^{(k)}\}_{k=1}^{K}$ and compute the variance of Sigmoid probabilities per pixel:

$$\mathbf{U}(i,j) = \mathrm{Var}_k\left[\sigma(\hat{\mathbf{x}}_0^{(k)}(i,j))\right], \tag{28}$$

which serves as the pixelwise uncertainty map. We average $\mathbf{U}$ within the mask and linearly rescale to obtain a slice-level quality score $q_{\mathrm{tcdr}} \in [0,1]$, used as a weight in subsequent volumetric consistency integration.

During training, only the diffusion refinement head is added and optimized with $\mathcal{L}_{\mathrm{TCDR}}$ until development convergence; at inference, conditioning on $\mathbf{P}_0$ and $\{\Phi(I), \mathbf{e}_t\}$, few-step DDIM sampling returns $\{\mathbf{P}_{\mathrm{ref}}, \hat{\mathbf{M}}_{\mathrm{ref}}, \mathbf{U}, q_{\mathrm{tcdr}}\}$. TCDR does not alter SEEM's input form, requires no points or boxes, and relies only on the terminology phrase and the raw slice for conditioning.

After obtaining slice-wise refined masks $\hat{\mathbf{M}}_{\mathrm{ref},t}$ and their quality scores $q_{\mathrm{tcdr},t}$ (from TCDR), we process the same volume as an axial sequence $\{I_t\}_{t=1}^{T}$. We propagate and correct segmentations between adjacent slices via STM, and maintain a limited number of candidate paths using a training-free trellis to avoid error cascading along a single path. Concretely, for slice t, constrained by the previous slice's final mask $\mathbf{M}_{t-1}^{\star}$ and the current refined probability map $\mathbf{P}_{\mathrm{ref},t}$, we generate up to K candidate masks $\{\tilde{\mathbf{M}}_t^{(k)}\}_{k=1}^{K}$ (including a propagation candidate that reuses the previous shape aligned to the current slice, and an evidence candidate obtained by thresholding $\mathbf{P}_{\mathrm{ref},t}$). For each candidate path we maintain an accumulated score

$$S_t^{(k)} = S_{t-1}^{(k)} + \lambda_1\,\mathrm{IoU}\big(\tilde{\mathbf{M}}_t^{(k)}, \mathbf{M}_{t-1}^{\star}\big) + \lambda_2\,\overline{P}\big(\tilde{\mathbf{M}}_t^{(k)}\big), \tag{29}$$

where IoU measures cross-slice overlap consistency and $\overline{P}$ is the mean confidence within the candidate mask (computed from $\mathbf{P}_{\mathrm{ref},t}$). This strategy uses only the two sources of evidence established earlier and inter slice consistency and TCDR pixel confidence Without introducing additional assumptions. At each slice we retain only the top-K paths for the next slice. In all experiments the STM module maintained at most $K = 3$ candidate paths per slice. Path scores used $\lambda_1 = 0.7$ for the overlap term and $\lambda_2 = 0.3$ for the confidence term, values that were selected on the HUG-ILD validation split by a small grid search that jointly considered Dice and HD95. Varying K between one and five or perturbing λ_1 and λ_2 by 0.2 changed Dice and HD95 by less than the observed variation across repeated runs, so we used this single configuration for all datasets.

After traversing the sequence, we select the full path with the highest accumulated score, $\{\mathbf{M}_t^\star\}_{t=1}^{T}$, as the volumetric segmentation result. Based on this path, we define a slice-level consistency score

$$c_t = \mathrm{norm}\big(\mathrm{IoU}(\mathbf{M}_t^\star, \mathbf{M}_{t-1}^\star),\ \overline{P}(\mathbf{M}_t^\star)\big) \in [0,1], \tag{30}$$

and fuse it with $q_{\mathrm{tcdr},t}$ to obtain voxel aggregation weights $w_t \propto c_t \cdot q_{\mathrm{tcdr},t}$ (normalized such that $\sum_t w_t = 1$), thereby down-weighting unstable slices in subsequent quantification. These two steps are aligned with our stated objectives: use sequence continuity to reinforce inter-slice coherence, and suppress long-range error propagation via an out-of-training trellis, without adding new modules or extra supervision.

For anatomy normalized quantification, we obtain bilateral lung and five-lobe masks $\{\mathbf{L}_{\mathrm{lung},t}, \mathbf{L}_t^{(\ell)}\}$ from an independent model decoupled from the present task; this model can be trained on or reused from public chest CT resources and challenges, lung and lobar segmentation benchmarks) to ensure anatomical consistency and cross-center reproducibility. Longstanding public challenges provide objective evaluation contexts for lung and lobe segmentation and can serve as one reference source for our independent anatomical segmenter. Let the per-voxel volume be v_{vox}. We compute three classes of volume-level metrics via voxelwise weighted aggregation:

(1) Overall fibrosis burden:

$$\eta \triangleq \%\mathrm{Fib} \ = \ \frac{\sum_t w_t \sum_{(x,y)\in\Omega_{\mathrm{lung},t}} \mathbf{1}\{\cdot\}[M_t(x,y)=1]\ v_{\mathrm{vox}}}{\sum_t w_t \left|\Omega_{\mathrm{lung},t}\right| v_{\mathrm{vox}}} \times 100\%\,. \tag{31}$$

(2) Lobar and side distribution:

$$\pi_\ell = \frac{\sum_t w_t \left|\mathbf{M}_t^\star \cap \mathbf{L}_t^{(\ell)}\right|}{\sum_t w_t \left|\mathbf{M}_t^\star \cap \mathbf{L}_{\mathrm{lung},t}\right|}, \qquad \pi_{\mathrm{R/L}} = \sum_{\ell\in\mathrm{R/L}} \pi_\ell. \tag{32}$$

(3) Subpleural preference and distance percentiles: within the lungs, compute a normalized pleural distance $D_t(x) \in [0,1]$ (0 at the pleura, 1 toward the hilum), and operationalize the subpleural region by the outer band $\mathcal{B}_t(\alpha) = \{x \in \mathbf{L}_{\mathrm{lung},t} \mid D_t(x) \le \alpha\}$ (with α fixed on the development set). Then,

$$\rho_{\mathrm{subpleural}} = \frac{\sum_t w_t \left|\mathbf{M}_t^\star \cap \mathcal{B}_t(\alpha)\right|}{\sum_t w_t \left|\mathbf{M}_t^\star \cap \mathbf{L}_{\mathrm{lung},t}\right|}, \qquad Q_p = \mathrm{Quantile}_p\big(\{D_t(x) \mid x \in \mathbf{M}_t^\star\}\big). \tag{33}$$

The imaging meaning and usage conventions of spatial terms such as subpleural and basal predominance can be cross-checked with the Fleischner glossary; translating them into quantitative definitions via distance thresholds and percentiles facilitates consistent review across centers and acquisition protocols.

Finally, we serialize %-fibrosis, $\{\pi_\ell\}$, $\rho_{\mathrm{subpleural}}$, $\{Q_{25}, Q_{50}, Q_{75}\}$, and the time series $\{c_t\}$ into a structured quantitative output, accompanied by template descriptions aligned with Fleischner terminology, for VILA-M3 to perform cross-modal consistency checks and auditable logging. The voxel-level annotations and clinical conventions provided by HUG-ILD furnish a reviewable data basis for the closed loop spanning terminology with mask alignment, TCDR refinement quality, and volume-level quantification. We report Q_{25}, Q_{50} and Q_{75} as the empirical quartiles of this distance distribution within the final lesion mask for each study and keep this definition fixed for all cohorts so that the indices are comparable across scanners.

4.6 VILA-M3 Cross-modal Consistency Auditing

As the system's top-level controller, VILA-M3 coordinates the CXR language agent and the CT segmentation, a quantification agent through a standardized tool-calling framework, and enforces a structured cross-modal consistency audit to support clinical trustworthiness and traceability. The complete workflow and key technical steps are as follows.

VILA-M3 first triggers the CXR language agent's classification interface to obtain two core outputs: (i) the temperature-calibrated four-dimensional ILD finding probability vector $\mathbf{p}_{\mathrm{ILD}}$ corresponding to fibrosis and (ii) the set of template phrases triggered by high-confidence findings, $\mathcal{T}(\mathbf{p}_{\mathrm{ILD}})$. High-confidence findings are selected using decision thresholds $\{\theta_k\}$ optimized on the validation set for $k \in \{\mathrm{fib}, \mathrm{ret}, \mathrm{hon}, \mathrm{tbe}\}$, determined by the indicator

$$\mathbb{I}\{p_k \geq \theta_k\}. \tag{34}$$

Only when a finding meets its threshold is the corresponding template phrase included in $\mathcal{T}(\mathbf{p}_{\mathrm{ILD}})$. When a calibrated probability p_k falls below its threshold θ_k, the corresponding template phrase is not included in $T(p_{\mathrm{ILD}})$ and the control prefix carries only the numerical value p_k, so near threshold fluctuations change the strength of the soft bias but do not generate a categorical positive statement. VILA-M3 then normalizes the phrases in $\mathcal{T}(\mathbf{p}_{\mathrm{ILD}})$ according to the Fleischner Society Glossary, translating qualitative CXR findings into standard terminology prompts required by CT segmentation to ensure semantic consistency across modalities.

After processing CXR-side evidence, VILA-M3 invokes two core interfaces of the CT segmentation. First, the slice-level segmentation interface is called, feeding the normalized terminology prompts slice by slice to obtain, for each CT slice i ($i = 1, 2, \ldots, N$, where N is the number of slices), a binary lesion mask M_i and a slice-level quality score q_i. Here q_i is computed by averaging the pixelwise uncertainty map $\mathbf{U}_i$ (from TCDR) within M_i and linearly rescaling to $[0, 1]$, reflecting the reliability of the slice segmentation. Second, the series-level aggregation interface uses pre-computed whole-lung and lobar anatomical masks to spatially align and quantify $\{M_i\}$ in 3D, returning three structured metrics together with visualizations of lesion masks overlaid on CT: the whole-lung lesion volume percentage

$$\eta = \frac{\sum_{i=1}^{N} |M_i|}{\sum_{i=1}^{N} |L_i|} \times 100\%, \tag{35}$$

where L_i is the lung mask for slice i and $|\cdot|$ counts voxels; the lobar distribution bias

$$\Delta_{\text{lobe}} \;=\; \frac{\sum_{i=1}^{N} |M_{i,\text{low}}| - \sum_{i=1}^{N} |M_{i,\text{up}}|}{\sum_{i=1}^{N} |M_i|}, \tag{36}$$

where $M_{i,\text{low}}$ and $M_{i,\text{up}}$ denote lower- and upper-lobe lesion masks on slice i; and the subpleural lesion percentage

$$\gamma \;=\; \frac{\sum_{i=1}^{N} |M_{i,\text{sub}}|}{\sum_{i=1}^{N} |M_i|} \times 100\%, \tag{37}$$

where $M_{i,\text{sub}}$ denotes the lesion mask within the subpleural band on slice i.

For each study we reuse these indices to define term level audit deviations. Subpleural claims are mapped to the scalar index γ and lower lobe predominance claims are mapped to Δ_{lobe}, and in both cases we compute an absolute deviation between the normalized textual claim and the corresponding CT index. A claim is considered supported when this deviation is not greater than the global tolerance ϵ introduced above. For reporting we define Consistency@term by restricting to studies that contain at least one audited claim and marking a study as consistent when all of its audited claims are supported under this rule. Consistency@term is the proportion of such studies within the evaluation set.

With CXR-side structured report fragments, including finding descriptions and preliminary clinical conclusions) and CT-side quantitative metrics and visualizations in hand, VILA-M3 integrates multi-source information following the logic of clinical radiology reporting to produce a composite report containing narrative text, key quantitative data, and visual materials. To ensure cross-modal consistency, VILA-M3 simultaneously performs a standardized audit by constructing quantitative criteria

$$\Delta_{\text{audit}} \;=\; \left| s_{\text{text}} - s_{\text{CT}} \right| \;\leq\; \epsilon, \tag{38}$$

where s_{text} is a qualitative mapping of the textual statement, subpleural predominant mapped to 0.6, lower lobe predominant mapped to 0.6), s_{CT} is the normalized value of the corresponding CT metric, γ or $|\Delta_{\text{lobe}}|$), and ϵ is a tolerance. In our implementation we set $\epsilon = 0.18$, chosen on the HUG-ILD validation split by sweeping values between 0.10 and 0.25 and selecting the smallest value that kept at least ninety percent of cases judged consistent by an internal radiologist review while still flagging a meaningful fraction of mismatches. We repeated this sweep after stratifying cases by ILD subtype and by lower-lobe versus upper-lobe predominant involvement and found that the optimal ϵ varied by at most 0.02, so we adopted a single global tolerance rather than subtype-specific thresholds. If $\Delta_{\text{audit}} > \epsilon$, VILA-M3 highlights the discrepancy in the report and includes the numeric deviation for subsequent expert review or system-triggered secondary calibration.

Throughout the process, VILA-M3 enforces a fixed interface call order and standardized data exchange formats to ensure stable and reproducible module cooperation. All core intermediates for a given study, including p_{ILD}, the class thresholds, the selected phrase list $T(p_{\mathrm{ILD}})$, the control prefix and sentence level consistency scores, the slice masks $\{M_i\}$, the diffusion based uncertainty maps $\{U_i\}$ and quality scores $\{q_i\}$, the aggregate CT indices η, Δ_{lobe} and γ, and the term level audit decisions together with their deviations, are centrally stored by VILA-M3 as a single audit record. Optional artefacts such as rendered overlays and prototype trajectory plots are generated from this record and are not required to reconstruct the evidence chain.

4.7 Ethics approval and consent to participate

Not applicable. This work uses de-identified, publicly available datasets released under their respective data-use policies; no new human data were collected.

4.8 Consent for publication

Not applicable. This work exclusively utilizes de-identtified datasets available from public repositories.

Declarations

Data availability
All datasets used in this study are publicly accessible from the following official sources:MIMIC-CXR: https://physionet.org/content/mimic-cxr/2.0.0/
MIMIC-CXR-JPG (processed JPG version with standard splits): https://physionet.org/content/mimic-cxr-jpg/2.0.0/
HUG-ILD (HRCT with 3D annotations for interstitial lung disease): https://www.uhbs.ch/en/research/research-infrastructures/hug-ild-database
ReXGroundingCT (3D chest CT with text-finding to voxel mask alignments): https://arxiv.org/abs/2507.22030

Materials availability
Not applicable. No new physical materials were generated.

Code availability
All experiments were implemented in Python 3.10 using PyTorch (v2.3) with CUDA 12.1 and cuDNN 9, and were executed on four NVIDIA A100 GPUs with 80 GB memory each under a Linux environment. Medical image input output operations and sliding window inference follow MONAI (v1.5.1), and evaluation metrics are computed with TorchMetrics using synchronized reduction on a single device. Mixed precision training relies on the torch.amp autocast and GradScaler utilities, and all optimization, augmentation and calibration settings are exactly as specified in the Training details section to ensure reproducibility. The full training and inference code, together with configuration files and the random seeds used for all reported runs, will be publicly released on GitHub after formal publication of the paper.

Acknowledgement
None.This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

J.G., Y.R. and F.Y. contributed equally to this work, having full access to all study data and assuming responsibility for the integrity and accuracy of the analyses (Validation, Formal analysis). J.G. conceptualized the study, designed the methodology, and participated in securing research funding (Conceptualization, Methodology, Funding acquisition). Y.R. carried out data acquisition, curation, and investigation (Investigation, Data curation) and provided key resources, instruments, and technical support (Resources, Software). F.Y. drafted the initial manuscript and generated visualizations (Writing – Original Draft, Visualization). C.S., S.W., X.H. and C.C. supervised the project, coordinated collaborations, and ensured administrative support (Supervision, Project administration). All authors contributed to reviewing and revising the manuscript critically for important intellectual content (Writing – Review & Editing) and approved the final version for submission.

Conflict of interest/Competing interests

The author declares that there are no financial or non-financial competing interests relevant to the content of this work.

References

[1] Raghu, G., Remy-Jardin, M., Myers, J.L., Richeldi, L., Ryerson, C.J., Lederer, D.J., Behr, J., Cottin, V., Danoff, S.K., Morell, F., *et al.*: Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline. American journal of respiratory and critical care medicine **198**(5), 44–68 (2018)

[2] Chae, K.J., Hwang, H.J., Duarte Achcar, R., Cooley, J.C., Humphries, S.M., Kligerman, S., Lynch, D.A.: Central role of ct in management of pulmonary fibrosis. Radiographics **44**(6), 230165 (2024)

[3] Christensen, J.D., Harowicz, M., Walker, C.M., Little, B.P., Batra, K., Brixey, A.G., Carroll, M.B., Chelala, L., Cox, C.W., Drummond, M.B., *et al.*: Acr appropriateness criteria® chronic dyspnea-noncardiovascular origin: 2024 update. Journal of the American College of Radiology **22**(5), 163–176 (2025)

[4] Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J.: Fleischner society: glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (2008)

[5] Jacob, J., Bartholmai, B.J., Rajagopalan, S., Kokosi, M., Nair, A., Karwoski, R., Walsh, S.L., Wells, A.U., Hansell, D.M.: Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based ct analysis with conventional severity measures. European Respiratory Journal **49**(1) (2017)

[6] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)

[7] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., *et al.*: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15016–15027 (2023)

[8] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

[9] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

[10] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., *et al.*: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)

[11] Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature biomedical engineering **6**(12), 1399–1406 (2022)

[12] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in neural information processing systems **36**, 19769–19782 (2023)

[13] Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7086–7096 (2022)

[14] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082–18091 (2022)

[15] Ryu, J.S., Kang, H., Chu, Y., Yang, S.: Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomedical Engineering Letters, 1–22 (2025)

[16] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., *et al.*: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), 230024 (2023)

[17] Podolanczuk, A.J., Hunninghake, G.M., Wilson, K.C., Khor, Y.H., Kheir, F., Pang, B., Adegunsoye, A., Cararie, G., Corte, T.J., Flanagan, J., et al.: Approach

to the evaluation and management of interstitial lung abnormalities. an official american thoracic society clinical statement. American Journal of Respiratory and Critical Care Medicine (ja) (2025)

[18] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR

[19] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.*: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

[20] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

[21] Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)

[22] Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**(1), 29 (2015)

[23] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., *et al.*: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)

[24] Baharoon, M., Luo, L., Moritz, M., Kumar, A., Kim, S.E., Zhang, X., Zhu, M., Alabbad, M.H., Alhazmi, M.S., Mistry, N.P., et al.: Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. arXiv preprint arXiv:2507.22030 (2025)

[25] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[26] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[27] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[28] Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)

[29] Post, M.: A call for clarity in reporting bleu scores. arXiv preprint

arXiv:1804.08771 (2018)

[30] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72 (2005)

[31] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[32] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)

[33] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015). PMLR

[34] Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 375–383 (2017)

[35] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)

[36] Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)

[37] Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13753–13762 (2021)

[38] You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 72–82 (2021). Springer

[39] Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical image analysis **80**, 102510 (2022)

[40] Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11558–11567 (2023)

[41] Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19809–19818 (2023)

[42] Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)

[43] Bu, S., Song, Y., Li, T., Dai, Z.: Dynamic knowledge prompt for chest x-ray report generation. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 5425–5436 (2024)

[44] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv preprint arXiv:2004.09167 (2020)

[45] Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)

[46] Srivastav, S., Ranjit, M., Pérez-García, F., Bouzid, K., Bannur, S., Castro, D.C., Schwaighofer, A., Sharma, H., Ilse, M., Salvatelli, V., *et al.*: Maira at rrg24: A specialised large multimodal model for radiology report generation. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pp. 597–602 (2024)

[47] Tanno, R., Barrett, D.G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., *et al.*: Collaboration between clinicians and vision–language models in radiology report generation. Nature Medicine **31**(2), 599–608 (2025)

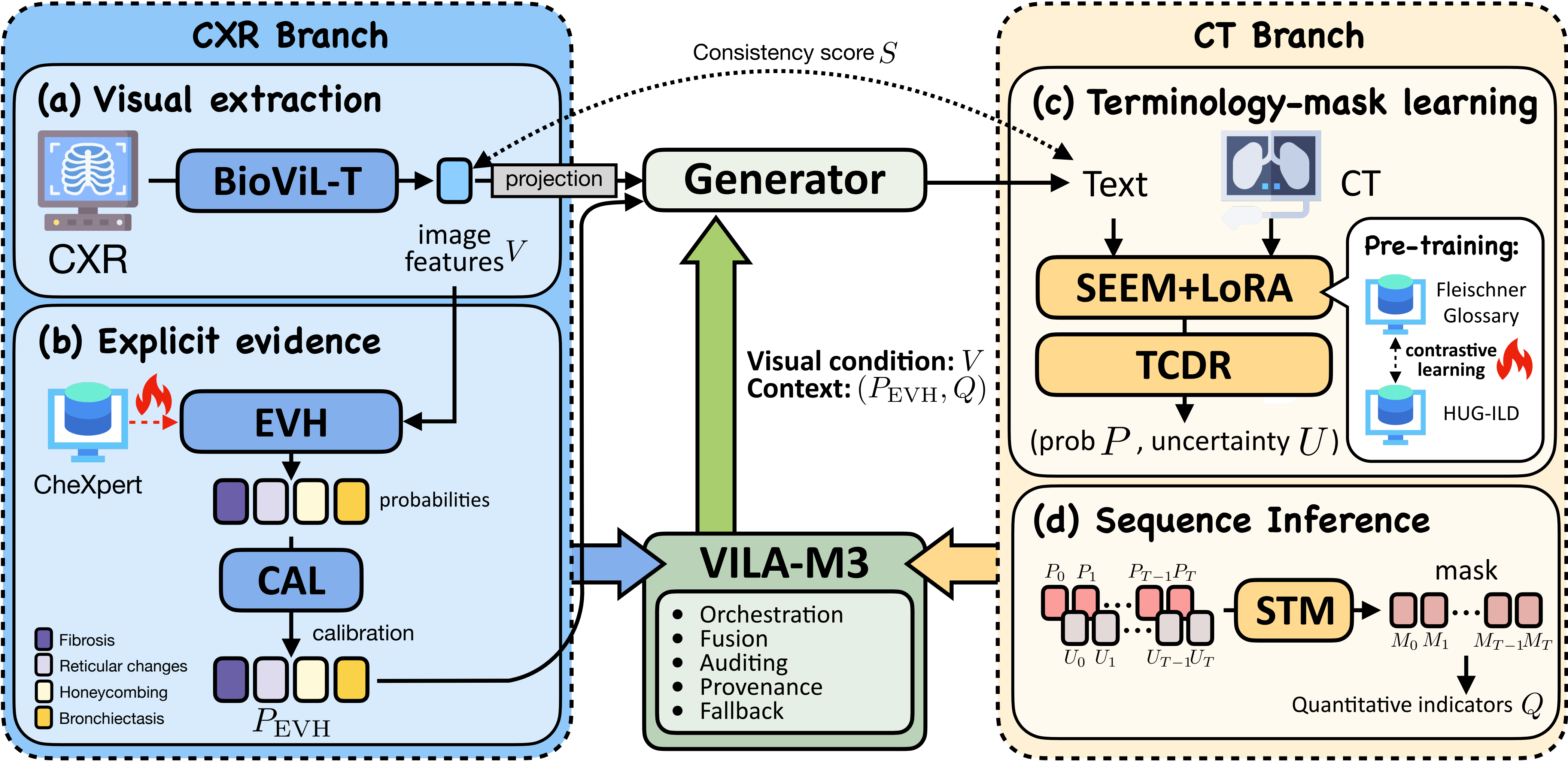
[48] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

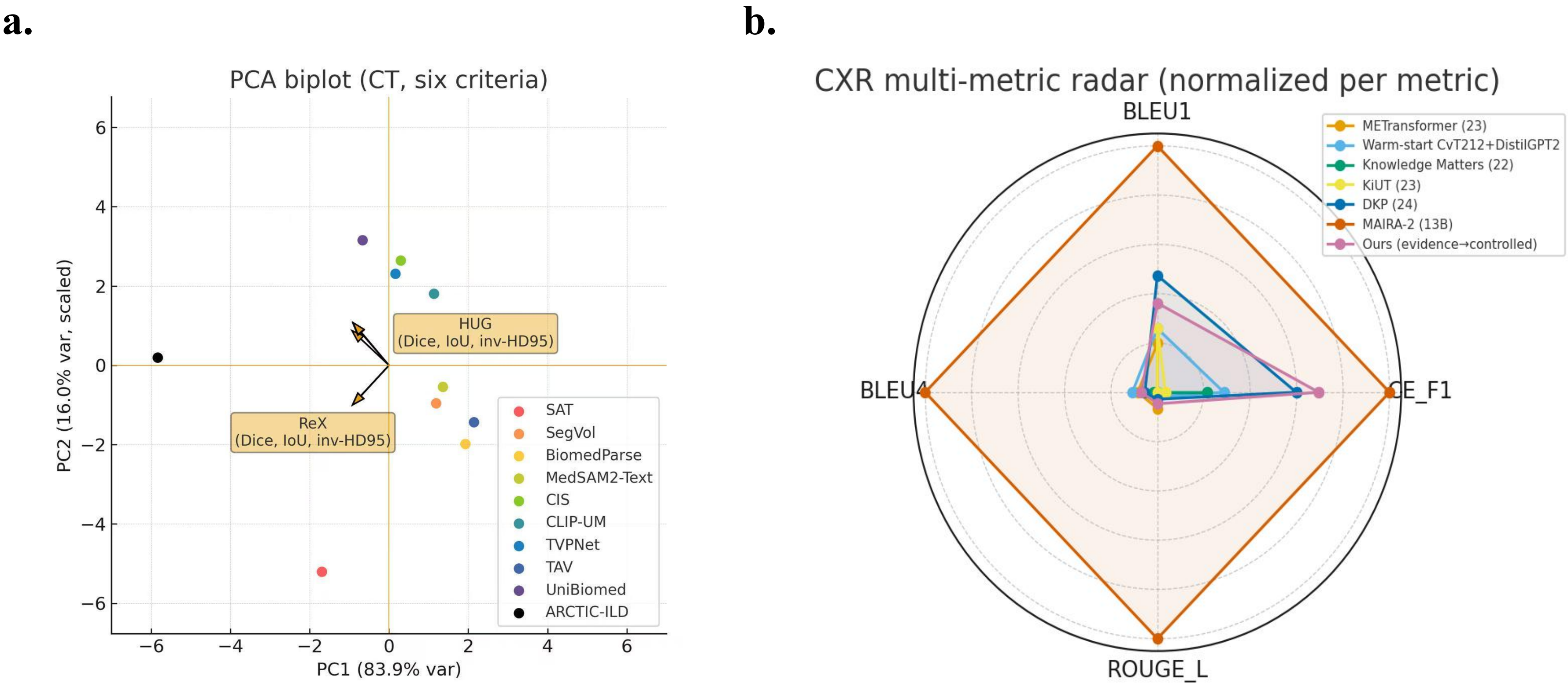
[49] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis, pp. 3–11 (2018). Springer

[50] Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)

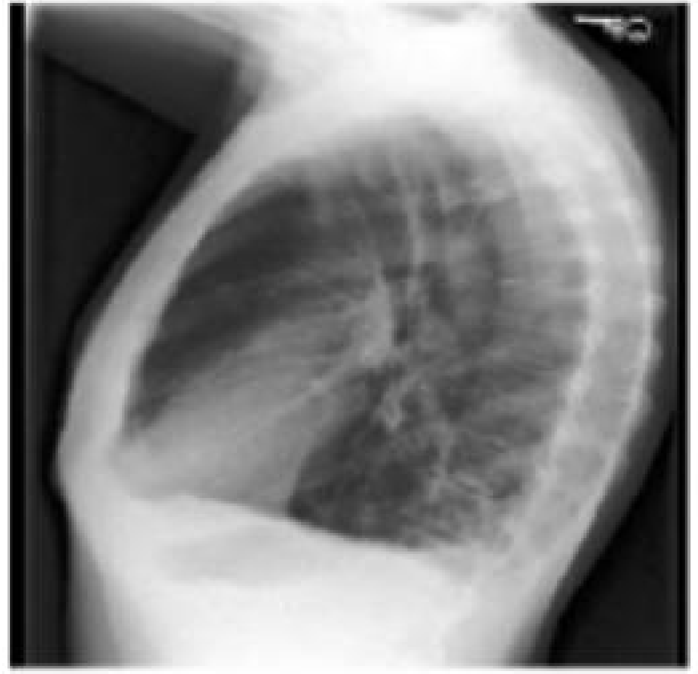
[51] Hatamizadeh, A., Yang, D., Roth, H., Xu, D.U.: Transformers for 3d medical image segmentation. arXiv preprint arXiv:2103.10504 (2021)

[52] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)


CXR Branch
(a) Visual extraction
CXR
BioViL-T
image features V
projection
Consistency score S
Generator
CT Branch
(c) Terminology-mask learning
Text
CT
SEEM+LoRA
TCDR
(prob P, uncertainty U)
Pre-training:
Fleischner Glossary
contrastive learning
HUG-ILD
Visual condition: V
Context: (P_{EVH}, Q)
(b) Explicit evidence
CheXpert
EVH
probabilities
CAL
calibration
P_{EVH}
Fibrosis
Reticular changes
Honeycombing
Bronchiectasis
VILA-M3
Orchestration
Fusion
Auditing
Provenance
Fallback
(d) Sequence Inference
P_0 P_1 P_{T-1} P_T
U_0 U_1 U_{T-1} U_T
STM
mask
M_0 M_1 M_{T-1} M_T
Quantitative indicators Q


a.
PCA biplot (CT, six criteria)
PC2 (16.0% var, scaled)
PC1 (83.9% var)
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
HUG
(Dice, IoU, inv-HD95)
ReX
(Dice, IoU, inv-HD95)
SAT
SegVol
BiomedParse
MedSAM2-Text
CIS
CLIP-UM
TVPNet
TAV
UniBiomed
ARCTIC-ILD
b.
CXR multi-metric radar (normalized per metric)
BLEU1
BLEU4
CE_F1
ROUGE_L
METransformer (23)
Warm-start CvT212+DistilGPT2
Knowledge Matters (22)
KiUT (23)
DKP (24)
MAIRA-2 (13B)
Ours (evidence→controlled)

Original image



pa and lateral chest radiographs were provided . the **cardiomediastinal silhouetteis** is unremarkable . there is a subtle opacityin the right lower lobe that is concerningfor **early pneumonia** . there is linearscarring in the left upper lobe from area of **prior pneumonia** that has resolved . thelungs are **hyperinflated** and the diaphragmsare **flattened consistent with copd** . there is no **pleural effusion** or **pneumothorax** there are no acute osseous lesions

GPT5-Thinking: *Key Findings*

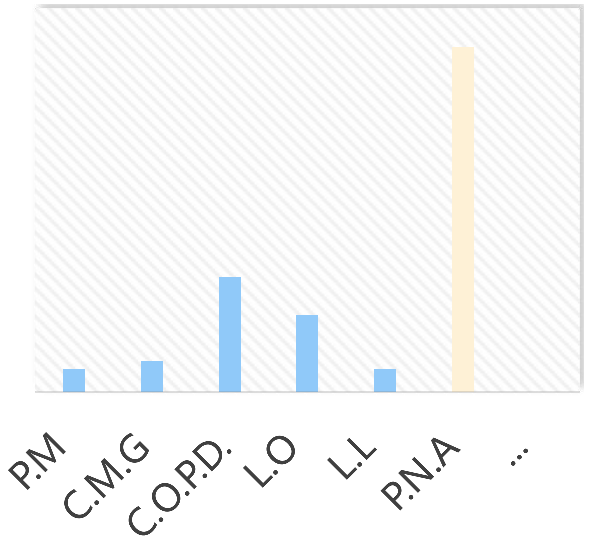
- *Ribs and thoracic vertebrae forming the posterior chest wall;*
- *Major **lung fields** without an obvious focal consolidation or large **pleural effusion** on this small image...*

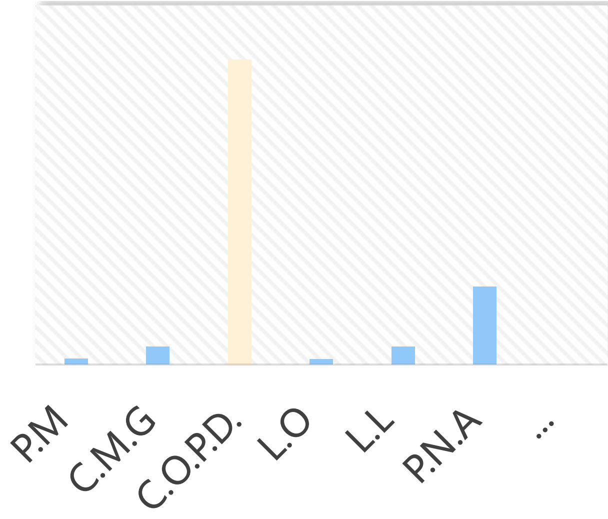
BioViL-T based baseline (Fine-tuned): *(Lateral chest radiograph: Ribs, thoracic vertebrae (darker inferiorly normally); cardiac silhouette (aerated lung behind);diaphragms/costophrenic angles (sharp normally, blunting = small effusion); major **lung fields** (no obvious consolidation/**effusion**);*

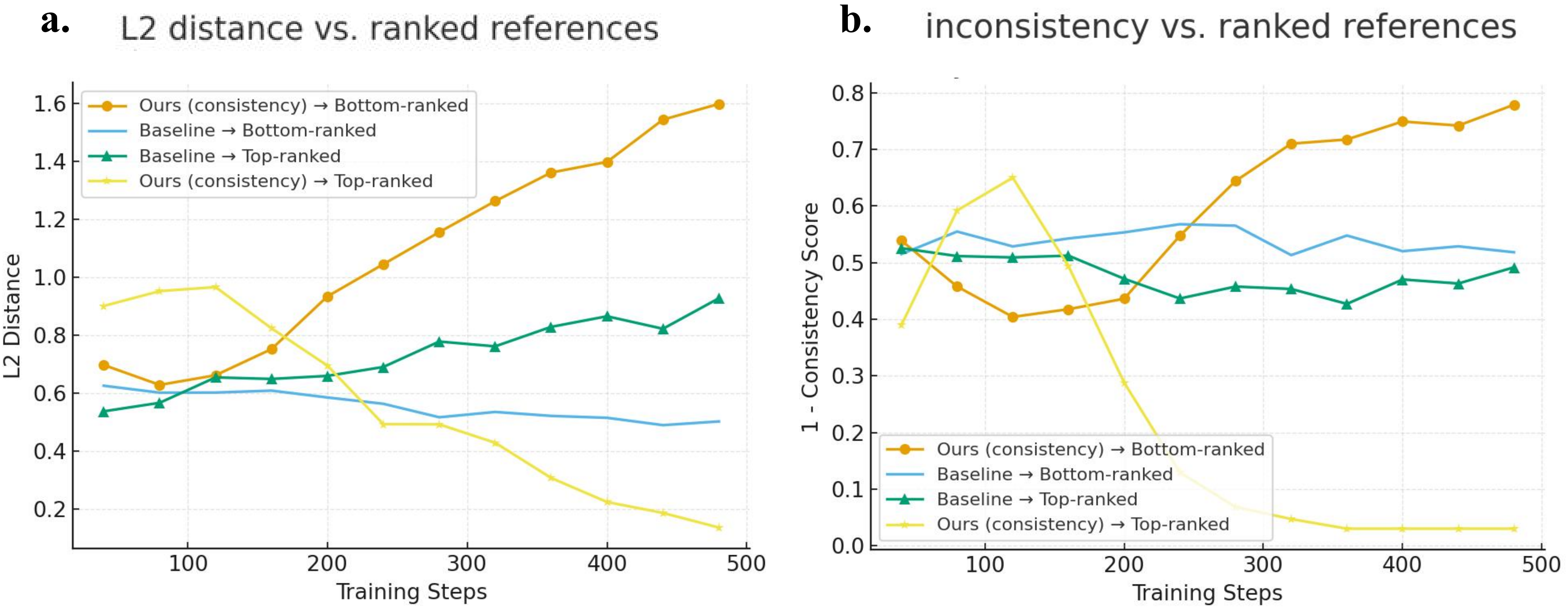
LLAVA-Med: *cardiomediastinal silhouette unremarkable. Subtle opacity in the left lower lobe suggesting early pneumonia. **Lungs hyperinflated** with **flattened diaphragms consistent with COPD**. Trace **pleural effusion**, no **pneumothorax**, osseous lesions.*

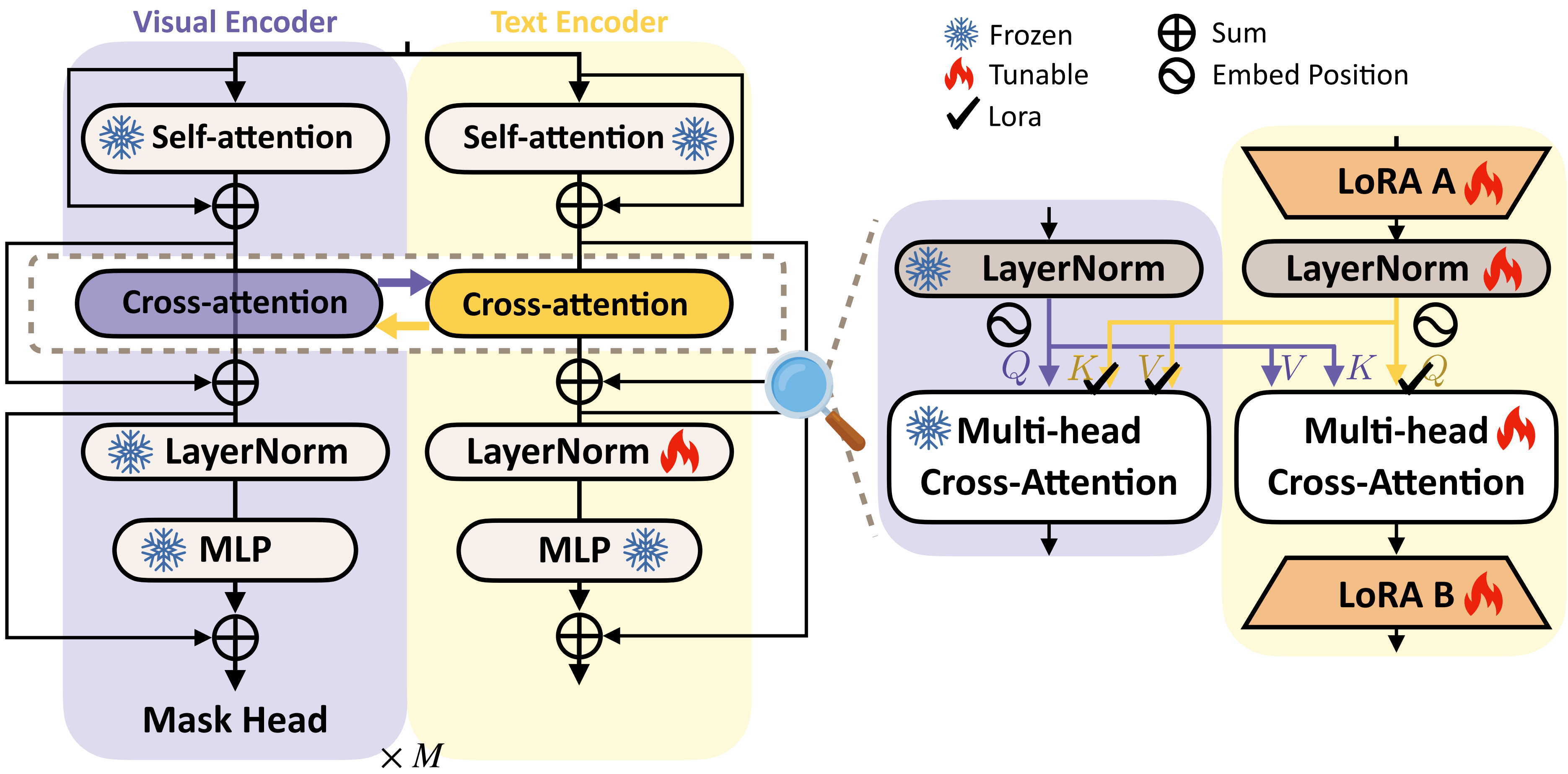
Med-PaLM M: Subtle right-lower-lobe opacity concerning for **early pneumonia**. Linear scarring in the right upper lobe from **prior infection**. **Lungs hyperinflated** with flattened diaphragms (**consistent with COPD**). No pleural effusion, no **pneumothorax**,.

cardiomediastinal silhouette unremarkable. Subtle right-lower-lobe opacity concerning for **early pneumonia**, and no evidence of **pneumothorax**. **Heart** size is mildly enlarged. Linear scarring in the left lobe from **prior pneumonia** (resolved). Lungs **hyperinflated** with **flattened diaphragms consistent with COPD**. No **pleural effusion**, no acute osseous abnormalities is present.






a.
L2 distance vs. ranked references
L2 Distance
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
100
200
300
400
500
Training Steps
Ours (consistency) → Bottom-ranked
Baseline → Bottom-ranked
Baseline → Top-ranked
Ours (consistency) → Top-ranked
b.
inconsistency vs. ranked references
1 - Consistency Score
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100
200
300
400
500
Training Steps
Ours (consistency) → Bottom-ranked
Baseline → Bottom-ranked
Baseline → Top-ranked
Ours (consistency) → Top-ranked


Visual Encoder
Text Encoder
Frozen
Tunable
Lora
Sum
Embed Position
Self-attention
Self-attention
Cross-attention
Cross-attention
LayerNorm
LayerNorm
MLP
MLP
Mask Head
$\times M$
LoRA A
LayerNorm
LayerNorm
Q
K
V
V
K
Q
Multi-head Cross-Attention
Multi-head Cross-Attention
LoRA B