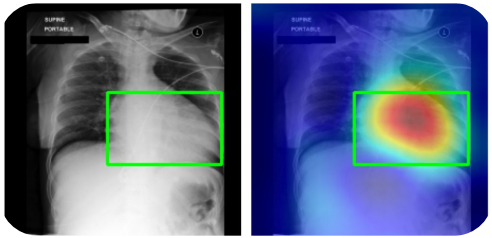
Narrative-Aware Domain Knowledge Integration for Annotation-Free Medical Image Localisation



Query: Bilateral, diffuse lung consolidation is present, consistent with acute lung injury.


Tahsir Ahmed Munna


Nuno Ricardo Guimarães


Alípio Mário Jorge

INESC TEC & FCUP | Porto, Portugal

tahsir.a.munna@inesctec.pt

ECIR 2026

ECIR@Text2Story'26 · Position Paper · Delft, 29 March 2026

INESCTEC

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

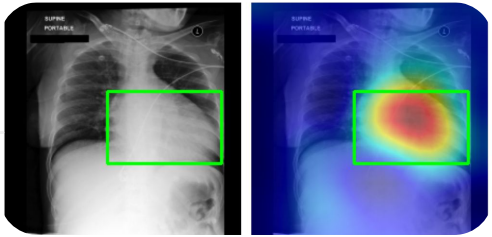
Key Term Unpacked

Medical Image Localisation?

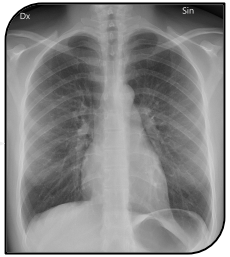
Vision-Language Model (VLM)?

Annotation Free?

Narrative-Aware?



Query: Bilateral, diffuse lung consolidation is present, consistent with acute lung injury.



Cardiomegaly

LOOKS

Enlarged cardiac silhouette (CT ratio > 0.5)

WHERE

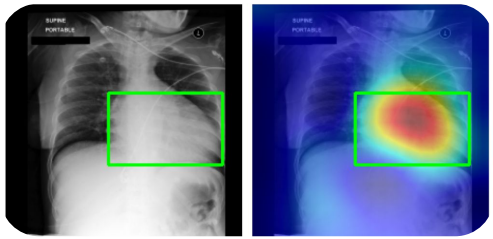
Central mediastinum

Millions of Images, But Who Draws the Boxes?

Localising pathological regions in chest X-rays is critical for diagnosis, monitoring, and model interpretability.

Traditional approach: *pixel-level annotations by expert radiologists.*

This is expensive, slow, and impossible to scale.



Query: Bilateral, diffuse lung consolidation is present, consistent with acute lung injury.

171K+

paired CXR–report examples

in MIMIC-CXR — yet spatial annotations cover only a fraction

Hours

of radiologist time per image

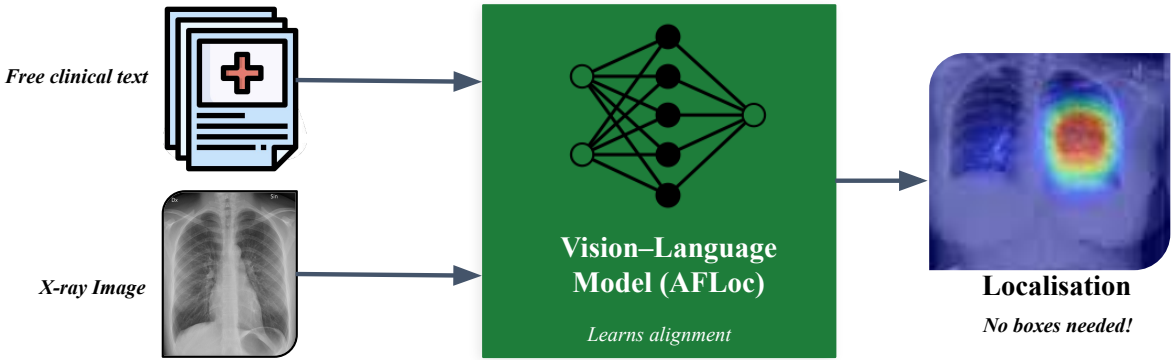
to produce reliable pixel-level ground truth at scale

High

inter-observer variability

even among experienced radiologists for the same case

Reports Are Useful — But Are They Enough?



✓ **Reports are generated in routine clinical workflows — annotation-free, no bounding boxes required, scalable**

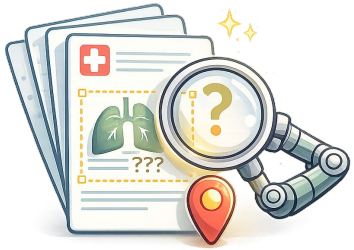
But this raises a fundamental question:

Are free-text radiology reports improve signal for spatial localisation of chest X-ray findings?

Reports Weren't Written for Machines to Localise

RADIOLOGY REPORT EXCERPT

*There is a small right pleural effusion that is **difficult to assess**. The cardiac silhouette is **somewhat mildly enlarged**.*



Actual clinical language

Why this limits localisation:

Stylistic variability

Phrasing differs across radiologists and institutions

Spatial vagueness

"Difficult to assess" ≠ a useful grounding signal

Hedged language

“Somewhat" — ambiguous for model training

Missing visual detail

Omits shape, density, extent — exactly what localisation needs

Implicit domain gap

Clinicians draw on background knowledge not stated in the text.

Our Position

Free-text reports are a starting point — not the full picture.

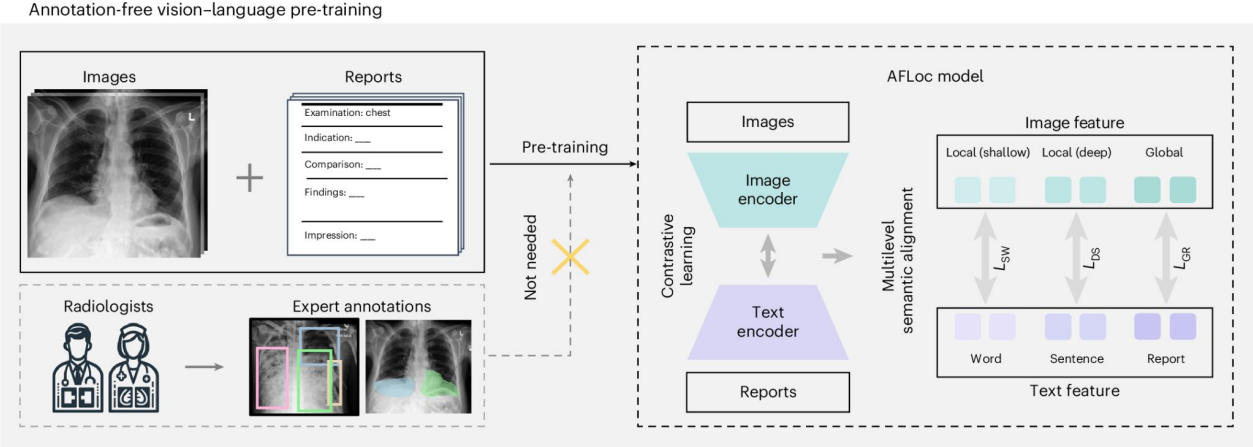
Enrich the text with structured domain knowledge, expressed in narrative form.

- Reports can be enhanced — not replaced
- Domain knowledge in narrative form adds spatial grounding
- Enriched supervision leads to better localisation
- Separating narrative from domain knowledge → clinical robustness

Text2Story theme: using narratives to enrich machines's vision

Baseline: AFLoc (*Yang et al., 2026)

A strong annotation-free VLM using paired CXR–report data from MIMIC-CXR as weak supervision.



Key: zero-pixel annotations — training uses only paired CXR images and free-text reports

We initialise from AFLoc's public checkpoint and fine-tune with our enriched supervision – no architectural changes.

**Yang, Hao, et al. "A multimodal vision–language model for generalizable annotation-free pathology localization." Nature Biomedical Engineering (2026): 1-15.*

Our Method: Narrative-Aware Knowledge Injection

BEFORE — Report Only

"There is a small right pleural effusion.

The cardiac silhouette is mildly enlarged."

Vague — no location, no appearance detail

The model must guess where to look.

inject
p=0.5

AFTER — Narrative-Enriched

[Original report]

"There is a small right pleural effusion. The cardiac silhouette is mildly enlarged."

[Disease description added]

Effusion: *"Homogeneous opacity with a meniscus sign blunting the costophrenic angle; dependent lower lateral lung zones tracking along the chest wall."*

Written by Expert

Clinically verified — not LLM-generated

Training-only

Not injected at inference

Fit Input Window

~40–45 tokens, leaves room for the report

Disease-Centric Narrative Descriptions (8 Pathologies)

Two-sentence template: Appearance → Location

Effusion

LOOKS

Homogeneous opacity with meniscus sign

WHERE

Lower lateral lung zones

Consolidation

LOOKS

Dense airspace opacity with air bronchograms

WHERE

Lower lobes, right lower lobe

Edema

LOOKS

Bilateral perihilar haziness, butterfly pattern

WHERE

Central hilar regions

Atelectasis

LOOKS

Linear opacities with volume loss

WHERE

Lower lobes, left base

Cardiomegaly

LOOKS

Enlarged cardiac silhouette (CT ratio > 0.5)

WHERE

Central mediastinum

Pneumothorax

LOOKS

Visible pleural line, absent lung markings

WHERE

Apex & upper lateral pleural space

Emphysema

LOOKS

Bilateral hyperlucency, flattened diaphragms

WHERE

Upper & middle lung zones

Pneumonia

LOOKS

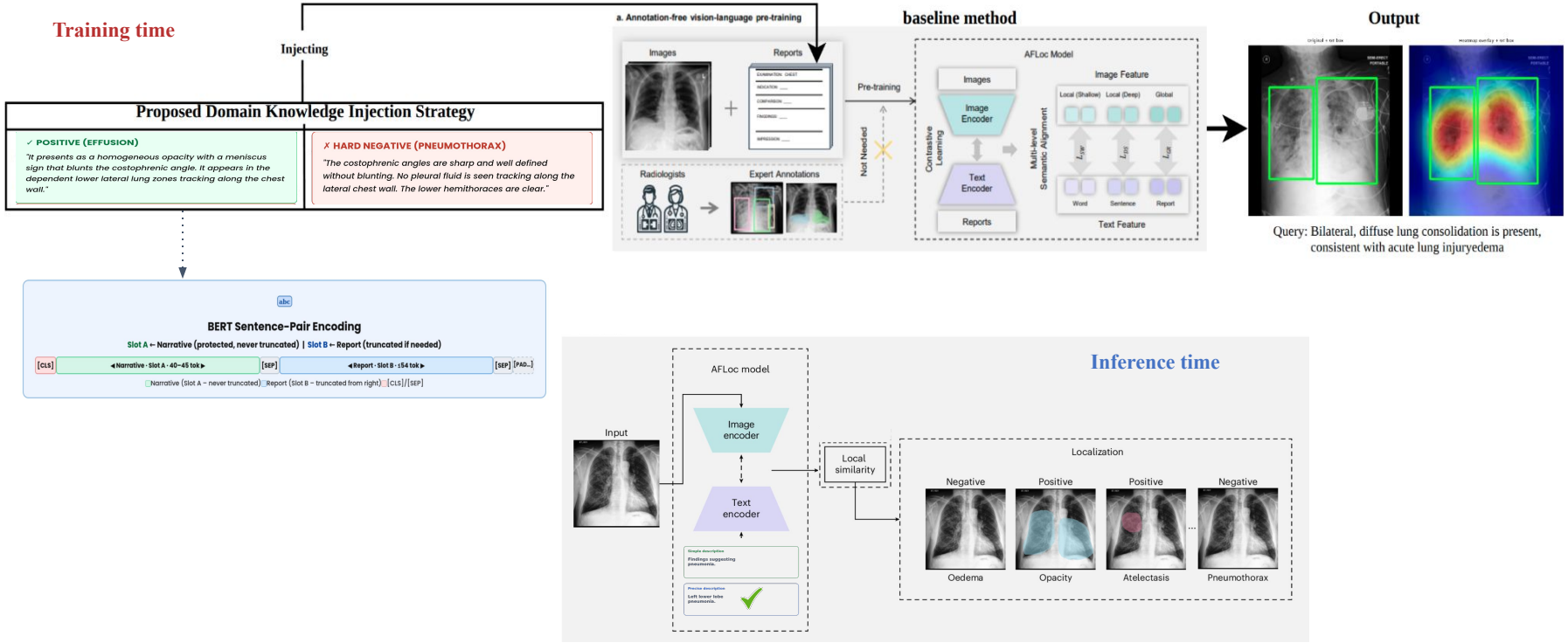
Segmental airspace opacity, indistinct borders

WHERE

Lower lobes, right middle lobe

Wu, Chaoyi, et al. "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis." *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.

Narrative Injection Pipeline



How We Measure: Metrics & Evaluation Protocol

MS-CXR benchmark · 1,162 chest X-ray–phrase pairs · 8 cardiopulmonary pathologies · Precise description setting

Three Localisation Metrics

IoU

Intersection over Union

$$\frac{|\text{Pred} \cap \text{GT}|}{|\text{Pred} \cup \text{GT}|}$$

Ratio of overlap to total area covered by predicted and ground-truth boxes. Range: 0 → 1.

Dice

Spatial overlap (harmonic mean)

$$\frac{2 \cdot |\text{Pred} \cap \text{GT}|}{|\text{Pred}| + |\text{GT}|}$$

Harmonic mean of precision and recall over the predicted region. Penalises false positives and negatives.

CNR

Contrast-to-Noise Ratio

$$(\mu_{region} - \mu_{bg}) / \sigma_{bg}$$

How distinctly the predicted activation region stands out from surrounding background tissue.

Results: Consistent Improvements Across All Metrics

MS-CXR benchmark · 1,162 CXR–phrase pairs · 8 pathologies · Precise description setting · Average across 5 thresholds

Method	mIoU	mDice	CNR
BioViL	0.228	0.342	1.083
GLoRIA	0.268	0.392	1.287
AFLoc (baseline)	0.324	0.462	1.636
Ours ✓	**0.330**	**0.470**	**1.671**

mIoU

0.324 → 0.330

+1.9%

Average Across 5 Thresholds

mDice

0.462 → 0.470

+1.7%

CNR

1.636 → 1.671

+2.1%

⚠ **Consistent, low-magnitude, directional evidence — not a final claim.**

Also outperforms BioViL & GLoRIA across all three metrics

Why It Works: Richer Text → Better Visual Grounding

No architectural change. No extra annotations. Just better semantic attributes.

Explicit Visual Attributes

Descriptions encode what models need: opacity type, density, and key radiographic signs.

An Honest Intermediate Step: Limitations & What's Next

Current Limitations

Modest gains not yet significant

More data and longer training needed to confirm

Only 8 pathologies — hand-crafting limits scale

Hard to extend to hundreds of conditions

Fixed injection probability (p = 0.5)

No adaptive selection — same rate for every sample

Future Directions

Narrative-Guided Contrastive Alignment Loss

Aligns visual patches with narrative-aware domain knowledge

Learned Injection Strategy

Replace fixed p=0.5 with adaptive, context-aware selection

Stronger Knowledge Separation

Disentangle narrative variability from domain knowledge

Larger-Scale Validation

Statistical evaluation to confirm gains and reduce variance

Key Takeaways

1

Free-text reports need enrichment, not replacement.

Scalable starting point — but not a sufficient supervision signal on their own.

2

Narrative-aware domain knowledge improves spatial grounding.

Consistent improvements in mIoU, mDice, and CNR — no architectural change, no extra annotation.

3

Knowledge expressiveness matters as much as availability.

How knowledge is represented — in narrative form — determines how well it guides visual learning.

Time To Answer The Question

Can structured narratives improve vision-language models to localise chest X-ray findings – without medical-image annotation?

Narrative-Aware Domain Knowledge Integration for Annotation-Free Medical Image Localisation

Thank you—Questions?


Tahsir Ahmed Munna


Nuno Ricardo Guimarães


Alípio Mário Jorge

INESC TEC & FCUP | Porto, Portugal


tahsir.a.munna@inesctec.pt