INNOPOLIS
UNIVERSITY

From Human Attention to Diagnosis: Semantic Patch-Level Integration of Vision-Language Models in Medical Imaging

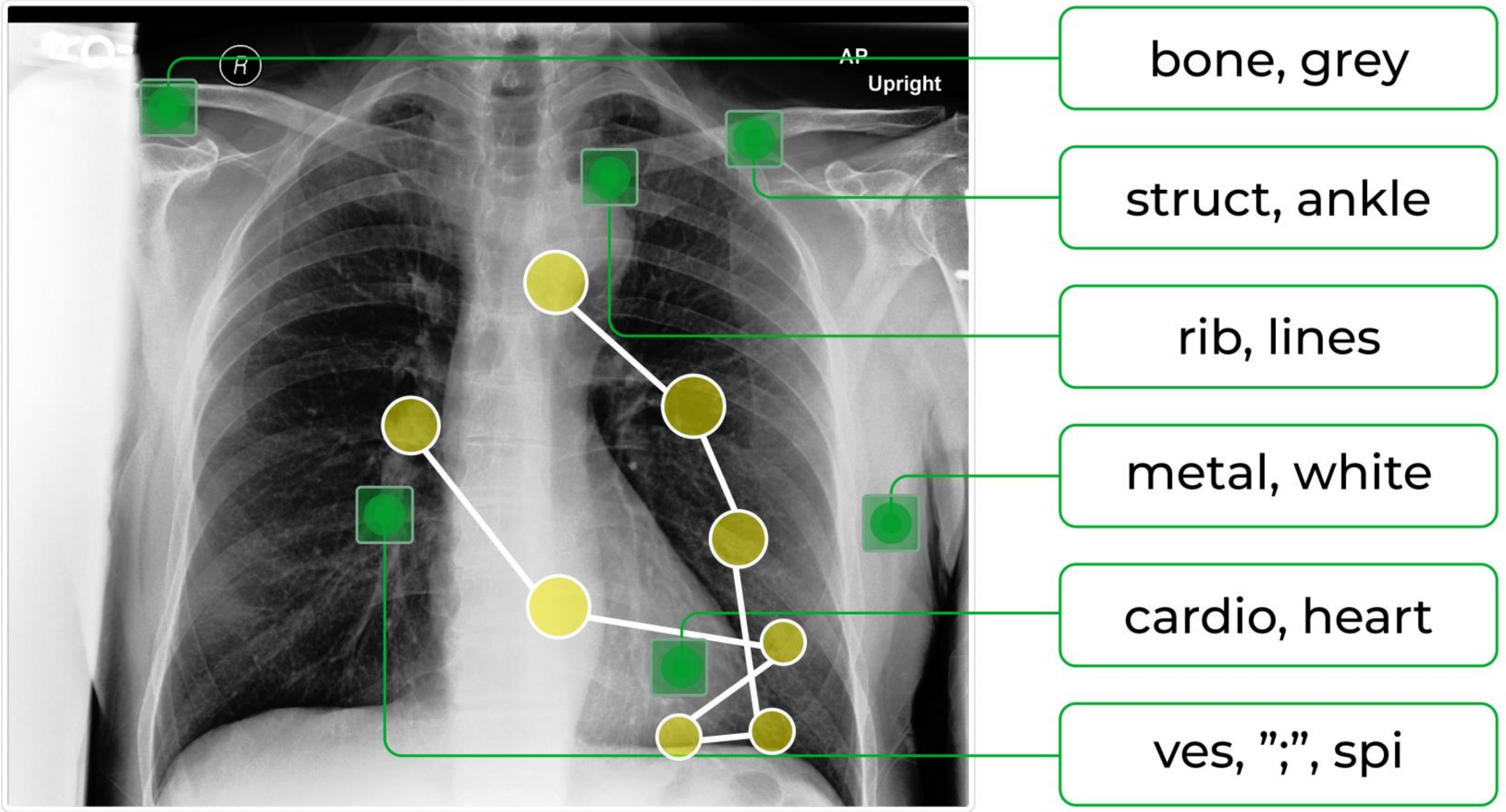
Dmitry Lvov, Ilya Pershin — Artificial Intelligence Institute, Innopolis University. Contact: d.lvov@innopolis.ru

Motivation

Why gaze?

- Expert eye movements encode diagnostic strategy.
- Fixations capture what clinicians consider important.
- We can use that signal to teach models **where** to look.

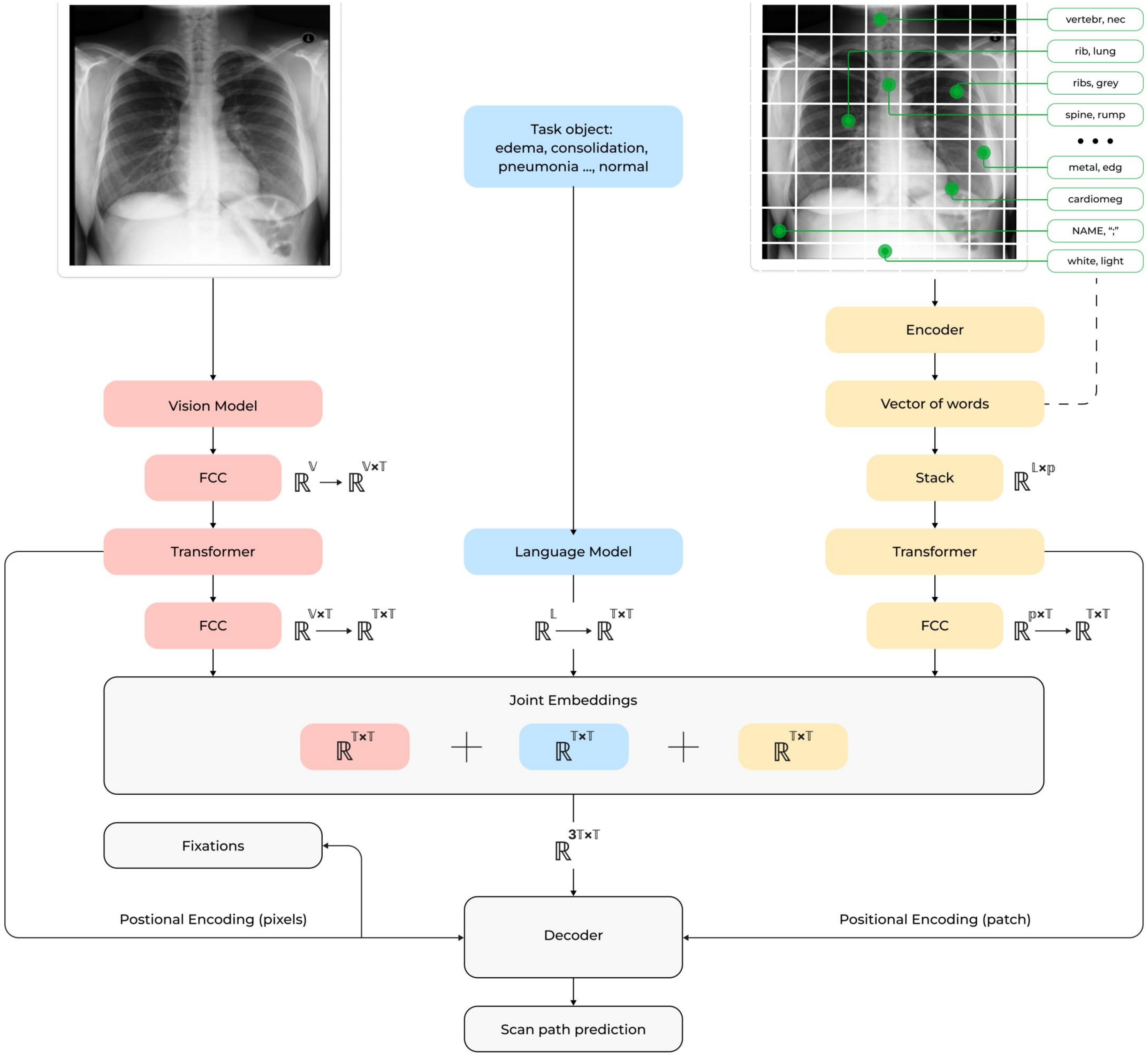
- Current gaze models mainly capture low-level saliency, ignoring patch-level meaning.
- We extract semantic signals from medical vision–language models and fuse them with visual features.
- Result: semantically-aware scanpaths that better reflect clinical reasoning.



Scanpath visualization on a chest X-ray

Method

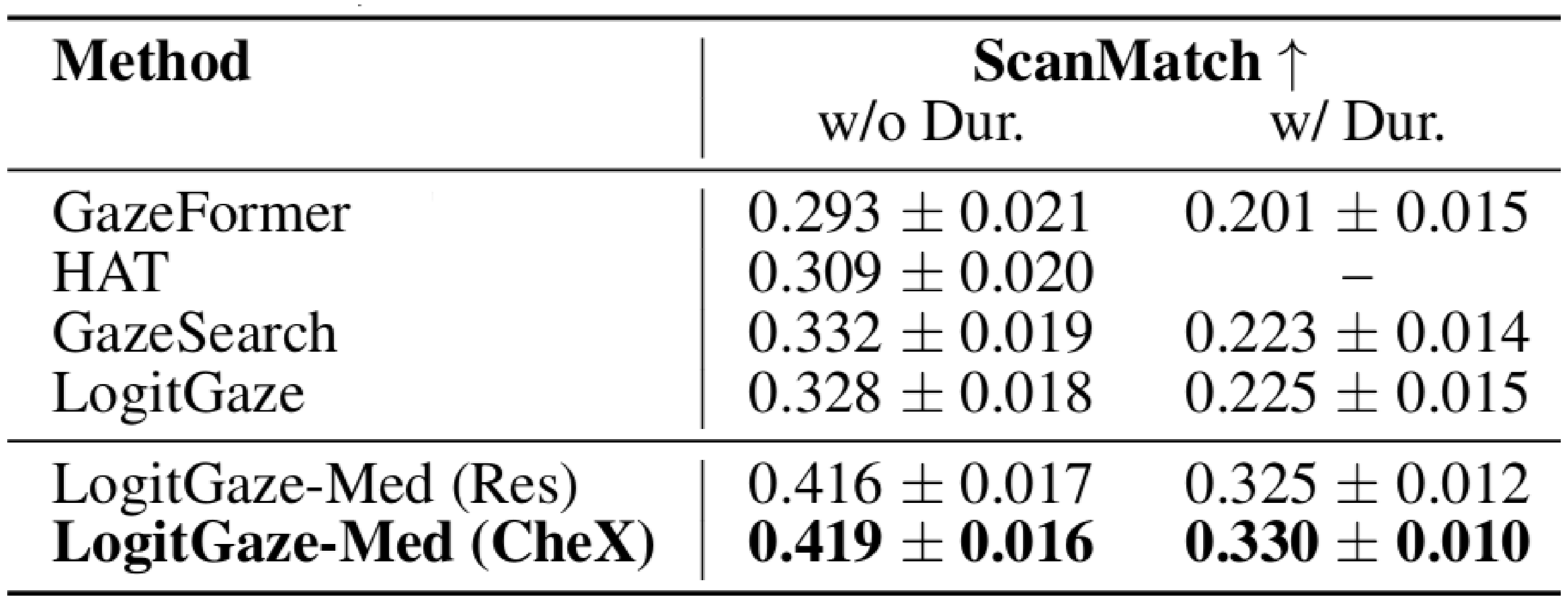
- Extract patch-level semantics from a medical VLM via a logit-lens.
- Fuse semantic vectors with visual features in a transformer.
- Predict continuous fixation coordinates (x,y) and dwell time per fixation.
- Sample stochastic scanpaths for downstream use.


vertebr, nec
rib, lung
ribs, grey
spine, rump
• • •
metal, edg
cardiomeg
NAME, “;”
white, light
Task object:
edema, consolidation,
pneumonia ..., normal
Vision Model
FCC
$\mathbb{R}^{V} \rightarrow \mathbb{R}^{V\times T}$
Transformer
FCC
$\mathbb{R}^{V\times T} \rightarrow \mathbb{R}^{T\times T}$
Language Model
$\mathbb{R}^{L} \rightarrow \mathbb{R}^{T\times T}$
Encoder
Vector of words
Stack
$\mathbb{R}^{L\times p}$
Transformer
FCC
$\mathbb{R}^{p\times T} \rightarrow \mathbb{R}^{T\times T}$
Joint Embeddings
$\mathbb{R}^{T\times T} + \mathbb{R}^{T\times T} + \mathbb{R}^{T\times T}$
$\mathbb{R}^{3T\times T}$
Fixations
Postional Encoding (pixels)
Decoder
Positional Encoding (patch)
Scan path prediction

LogitGaze-Med architecture

Results — realism and diagnostic benefit

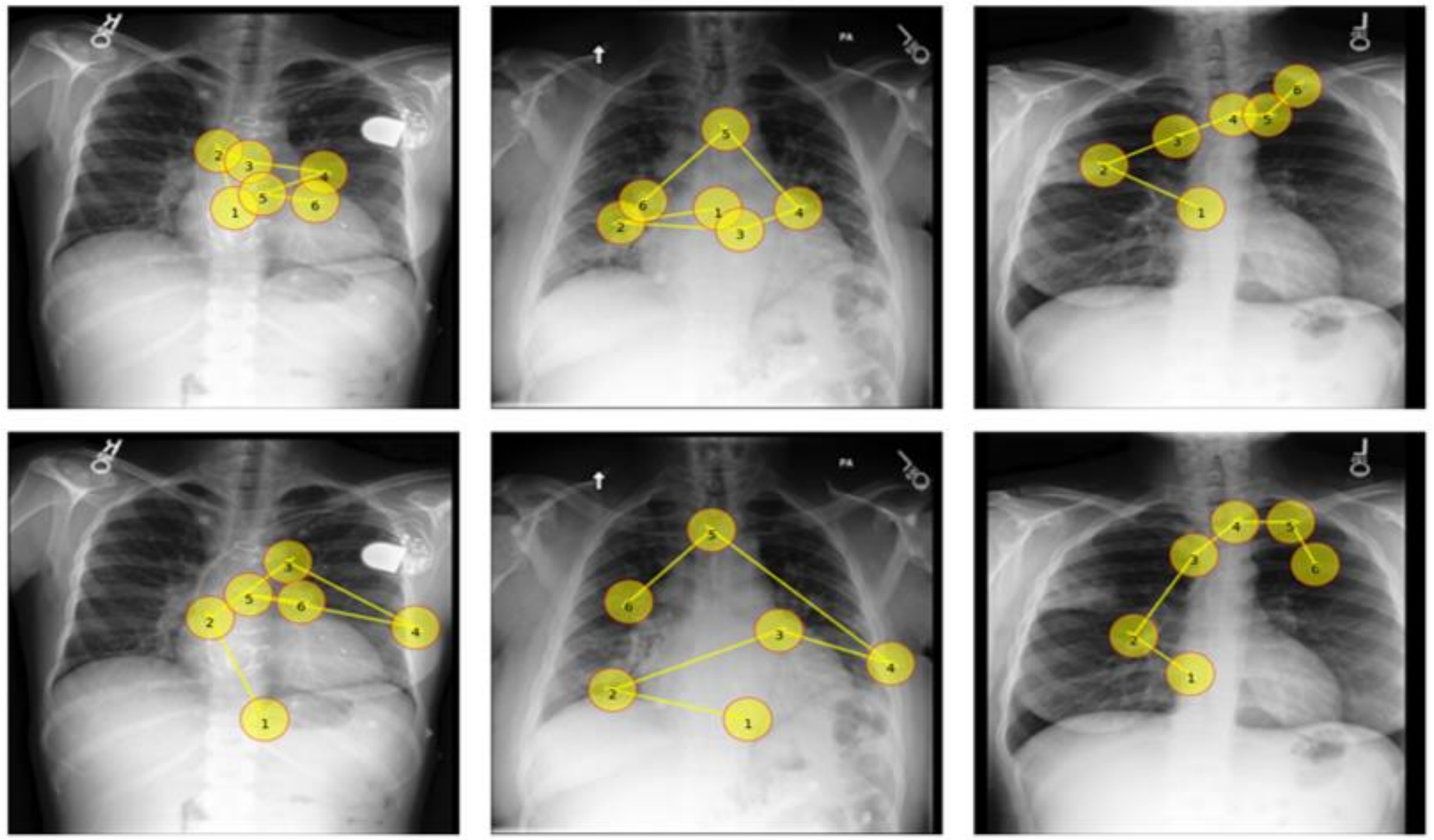
QUANTITATIVE



Method	ScanMatch ↑	
	w/o Dur.	w/ Dur.
GazeFormer	0.293 ± 0.021	0.201 ± 0.015
HAT	0.309 ± 0.020	–
GazeSearch	0.332 ± 0.019	0.223 ± 0.014
LogitGaze	0.328 ± 0.018	0.225 ± 0.015
LogitGaze-Med (Res)	0.416 ± 0.017	0.325 ± 0.012
LogitGaze-Med (CheX)	**0.419 ± 0.016**	**0.330 ± 0.010**

Performance on ScanMatch similarity metric

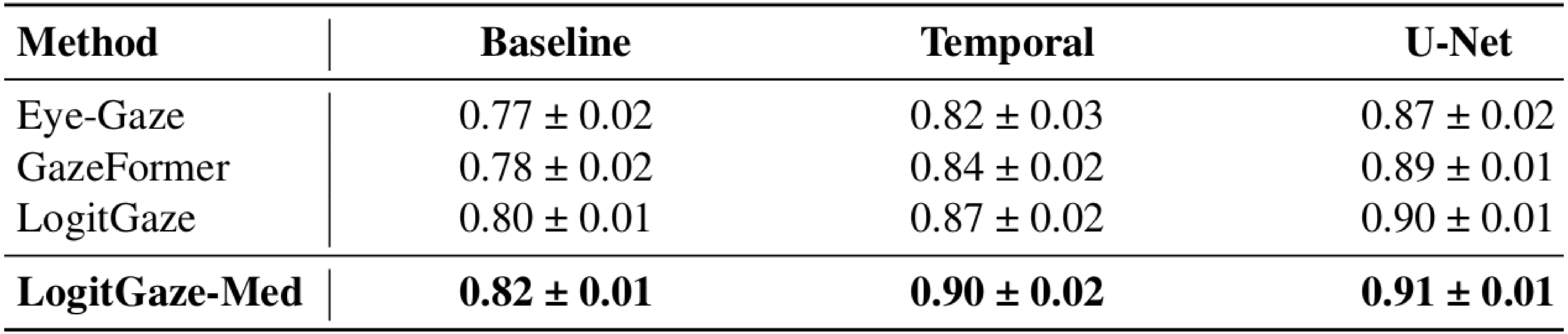
QUALITATIVE



Comparison of human scanpaths (top),
LogitGaze-Med predictions (bottom)

Results — realism and diagnostic benefit. Downstream task

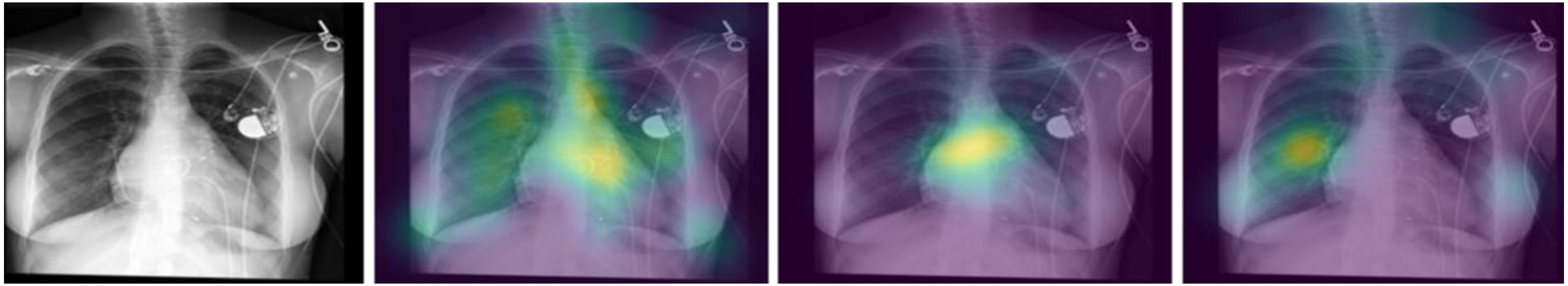
QUANTITATIVE



Method	Baseline	Temporal	U-Net
Eye-Gaze	0.77 ± 0.02	0.82 ± 0.03	0.87 ± 0.02
GazeFormer	0.78 ± 0.02	0.84 ± 0.02	0.89 ± 0.01
LogitGaze	0.80 ± 0.01	0.87 ± 0.02	0.90 ± 0.01
LogitGaze-Med	**0.82 ± 0.01**	**0.90 ± 0.02**	**0.91 ± 0.01**

AUROC scores across three classification setups

QUALITATIVE



(a) Original CXRs

(b) Human

(c) LogitGaze-Med

(d) Eye-Gaze baseline

Conclusions

Realism

Synthetic scanpaths closely match human patterns (↑ScanMatch 20–30%).

Effectiveness

Adding scanpaths improves diagnosis (AUROC +4–6 pp).

Practicality

Integrates with existing pipelines; requires more data and clinical validation.

Contact: d.lvov@innopolis.ru