# HYBRID EXPLANATION-GUIDED LEARNING FOR TRANSFORMER-BASED CHEST X-RAY DIAGNOSIS

**Shelley Zixin Shu**[1], Haozhe Luo[15], Alexander Poellinger[23], and Mauricio Reyes[14]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50, Bern 3008, Switzerland
[2] Inselspital (Bern University Hospital), 3010 Bern, Switzerland
[3] Insel Gruppe Bern Universitätsinstitut für Diagnostische, Interventionelle und Pädiatrische Radiologie
[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[5] Kaiko.AI, Zurich, Switzerland

# Background and Introduction

- Vision Transformers (ViTs) show **remarkable success** in medical image analysis.
- Transformer-based model empowered by the **attention mechanisms** also provides interpretability

However…..

- Machine features are not the same as human features in learning
- Prone to **shortcut learning**, **bias**, and **spurious correlations**

nature machine intelligence
PERSPECTIVE
https://doi.org/10.1038/s42256-020-00257-z
Check for updates

**Shortcut learning in deep neural networks**

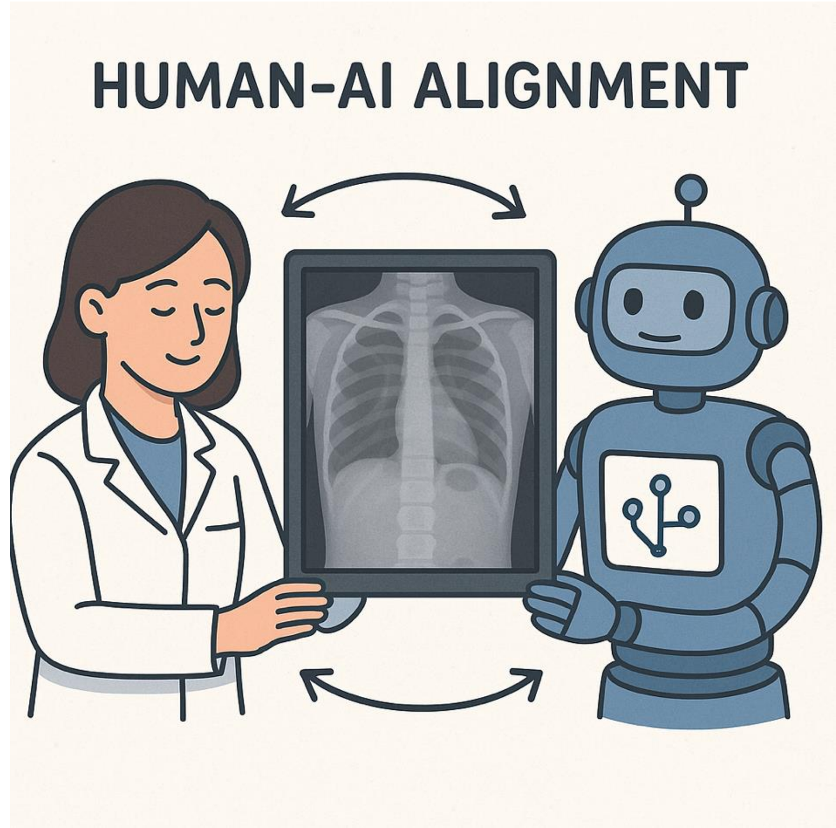
Robert Geirhos[1,2,4], Jörn-Henrik Jacobsen[3,4], Claudio Michaelis[1,2,4], Richard Zemel[3,5], Wieland Brendel[1,5], Matthias Bethge[1,5] and Felix A. Wichmann[1,5]

RESEARCH ARTICLE

Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

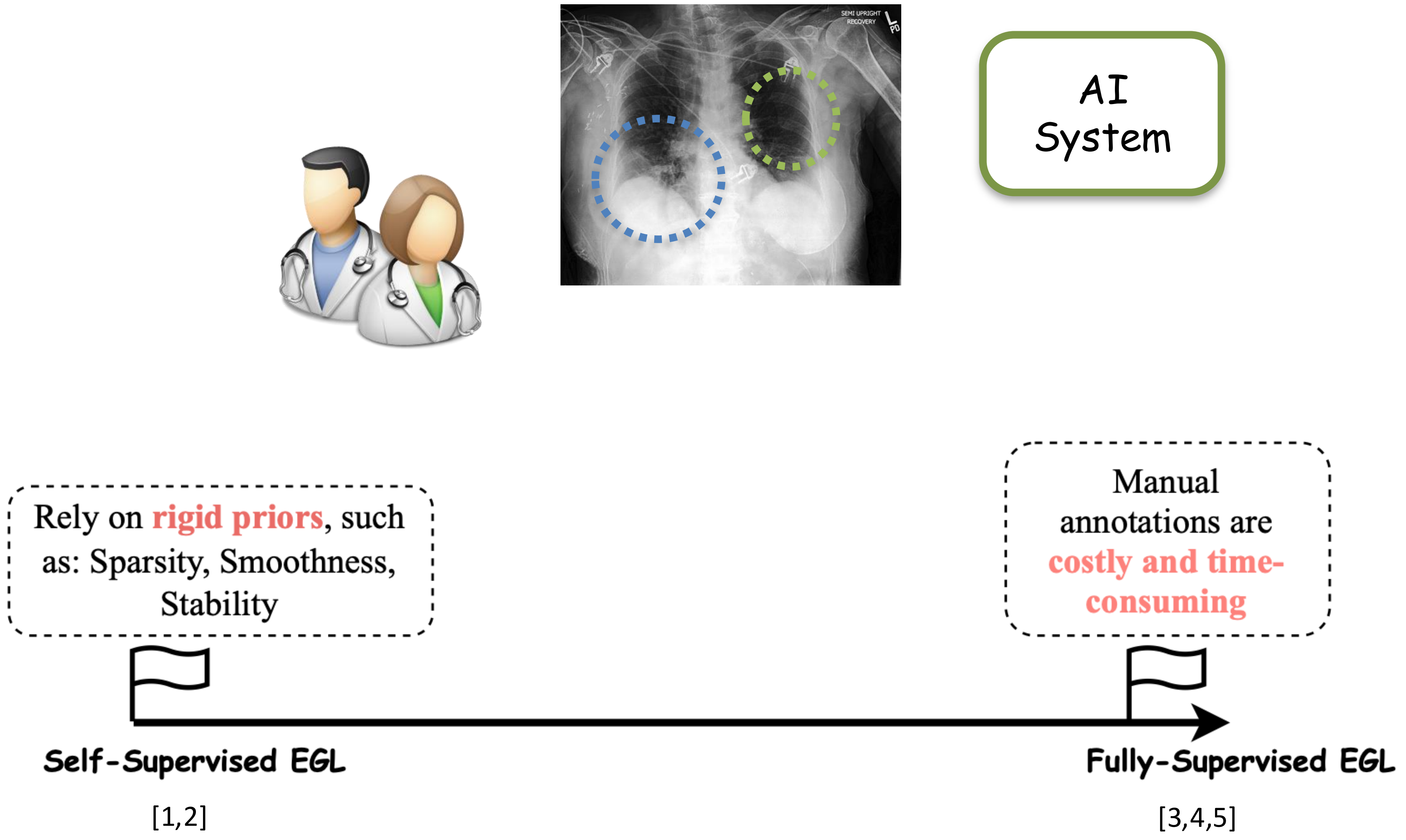
John R. Zech[1], Marcus A. Badgeley[2], Manway Liu[2], Anthony B. Costa[3], Joseph J. Titano[4], Eric Karl Oermann[3]*

1 Department of Medicine, California Pacific Medical Center, San Francisco, California, United States of America, 2 Verily Life Sciences, South San Francisco, California, United States of America, 3 Department of Neurological Surgery, Icahn School of Medicine, New York, New York, United States of America, 4 Department of Radiology, Icahn School of Medicine, New York, New York, United States of America



# Explanation-Guided Learning for Human-AI Alignment?

- Integrates **human knowledge** into training.
- Improves **robustness**, **fairness**, and **generalization** [6,7]
- Fully supervised approach use:
  - Expert-annotated explanations
  - Iterative human feedback



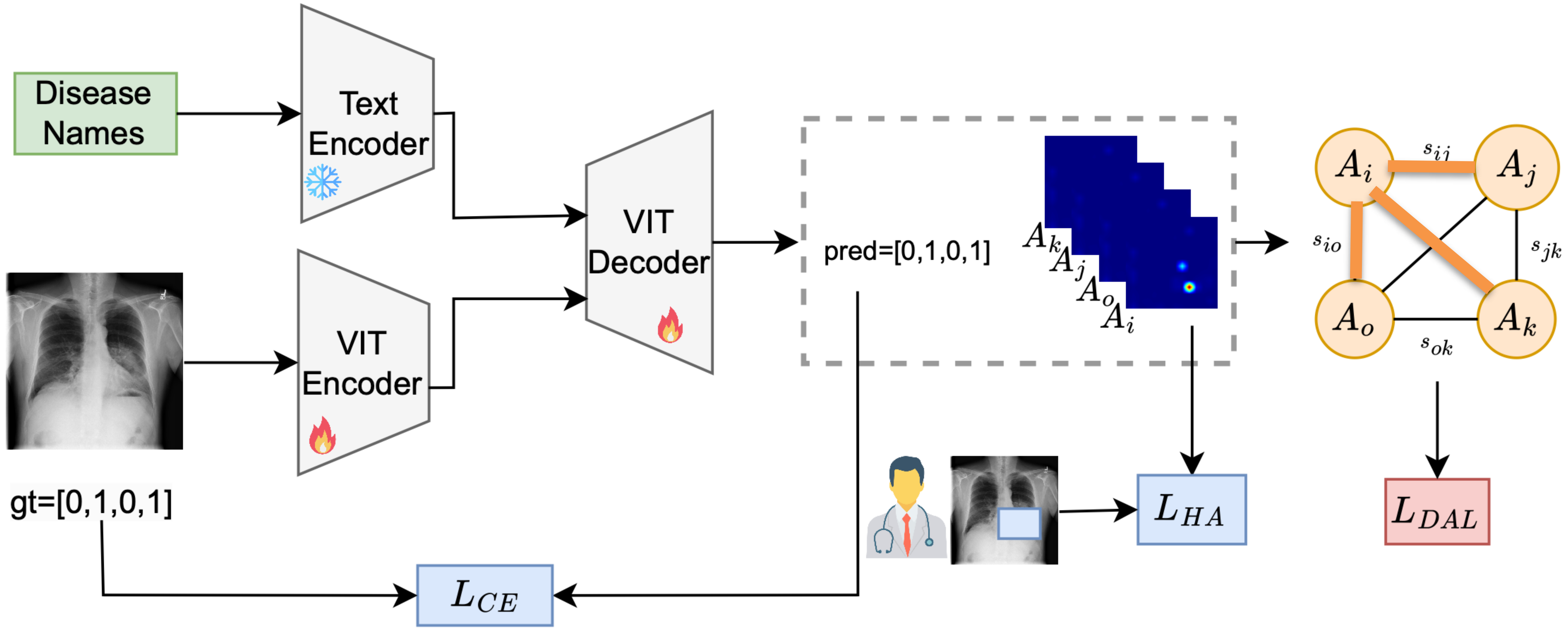

**Challenges:**

- **Manual annotations** are **costly and time-consuming**.
- Existing approaches, e.g. contrastive learning, self-supervised explanation learning, often rely on rigid priors (sparsity, smoothness and stability) → may **suppress complex clinical cues**.

We propose a **Hybrid Explanation-Guided Learning** (H-EGL) method to integrate self-supervised and expert-guided attention models for human-AI alignment.

[1] Singh, Krishna Kumar, et al. "Don't judge an object by its context: learning to overcome contextual bias." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[2] Pedapati, Tejaswini, et al. "Learning global transparent models consistent with local contrastive explanations." Advances in neural information processing systems 33 (2020): 3592-3602.
[3] Gao, Yuyang, et al. "Going beyond xai: A systematic survey for explanation-guided learning." ACM Computing Surveys 56.7 (2024): 1-39.
[4] Popordanoska, Teodora, Mohit Kumar, and Stefano Teso. "Machine guides, human supervises: Interactive learning with global explanations." arXiv preprint arXiv:2009.09723 (2020).
[5] Rieger, Laura, et al. "Interpretations are useful: penalizing explanations to align neural networks with prior knowledge." International conference on machine learning. PMLR, 2020.
[6] Rieger, L., Singh, C., Murdoch, W., Yu, B.: Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In: International conference on machine learning. pp. 8116–8126. PMLR (2020)
[7] Wu, S., Zhang, X., Wang, B., Jin, Z., Li, H., Feng, J.: Gaze-directed vision gnn for mitigating shortcut learning in medical image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 514–524. Springer (2024)

# Methodology




$$L_{\text{H-EGL}} = L_{CE} + \alpha L_{HA} + \beta L_{DAL}$$

$$\mathcal{L}_{\text{HA}} = 1 - \frac{2 \times |A_i \odot M_i|}{|A_i| + |M_i| + w_{FP} N_{FP}},$$

- $A_i$ is the model-generated attention map for class i
- $M_i$ is the corresponding expert mask
- $N_{FP}$ is the number of false positives
- $w_{FP}$ is a penalty coefficient.

$$\mathcal{L}_{\text{DAL}} = \frac{2}{C(C-1)} \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} \boxed{|S(A_i, A_j)|}$$

- C is the number of classes
- $S(A_i, A_j)$ denotes the cosine similarity between attention maps $A_i$ and $A_j$, as $s_{ij}$ on the above figure.

# Experiment Design

- **ChestXDet**, a subset of NIH ChestX-ray14, was used for experiments.
- It includes expert-annotated pathology bounding boxes for improved supervision
- Total of 3,578 patients: 3,025 in the training set, **553 in the test set**
- The training set is split into 80/20 train-validation using different random seeds
    - **Five** training runs were conducted for robustness evaluation
    - Results **averaged** across the five runs
- Evaluation performed on the **official test set** with AUC, F1, MCC and the generalisation gap between validation and test set.

# H-EGL helps improve accuracy and generalisation of the model

- **H-EGL achieves the best overall performance** across all evaluation metrics
- **Ablation study confirms** the strong impact of the **self-supervised component**
- **Smaller generalization gap** observed between validation and test sets, indicating better robustness

| | $\mathrm{AUC}_{test}\uparrow$ | $\mathrm{AUC}_{gap}\downarrow$ | $\mathrm{F1}_{test}\uparrow$ | $\mathrm{F1}_{gap}\downarrow$ | $\mathrm{MCC}_{test}\uparrow$ | $\mathrm{MCC}_{gap}\downarrow$ |
|---|---|---|---|---|---|---|
| KAD [19] | 88.1±0.3% | 2.5% | 68.2±2.5% | 1.8% | 57.5±2.3% | 4.8% |
| GAIN [9] | 88.0±0.4% | 2.7% | 67.8±2.2% | 2.4% | 57.2±2.0% | 5.6% |
| H-EGL (Ours) | **89.3±0.7%** | 1.5% | **69.4±1.9%** | **0.5%** | **58.3±2.5%** | **3.8%** |
| H-EGL | | | | | | |
| w/o $\mathcal{L}_{\mathrm{HA}}$ $\alpha$=0 | **89.3±1.0%** | **1.4%** | 67.6±1.2% | 1.4% | 56.5±1.6% | 5.2% |
| w/o $\mathcal{L}_{\mathrm{DAL}}$ $\beta$=0 [11] | 88.4±0.2% | 2.5% | 66.9±1.2% | 3.2% | 56.3±1.0% | 6.5% |

*Baselines*

- *KAD* utilizes knowledge graphs for improved visual reasoning.
- *GAIN* enhances interpretability via attention guided by cross-entropy loss.

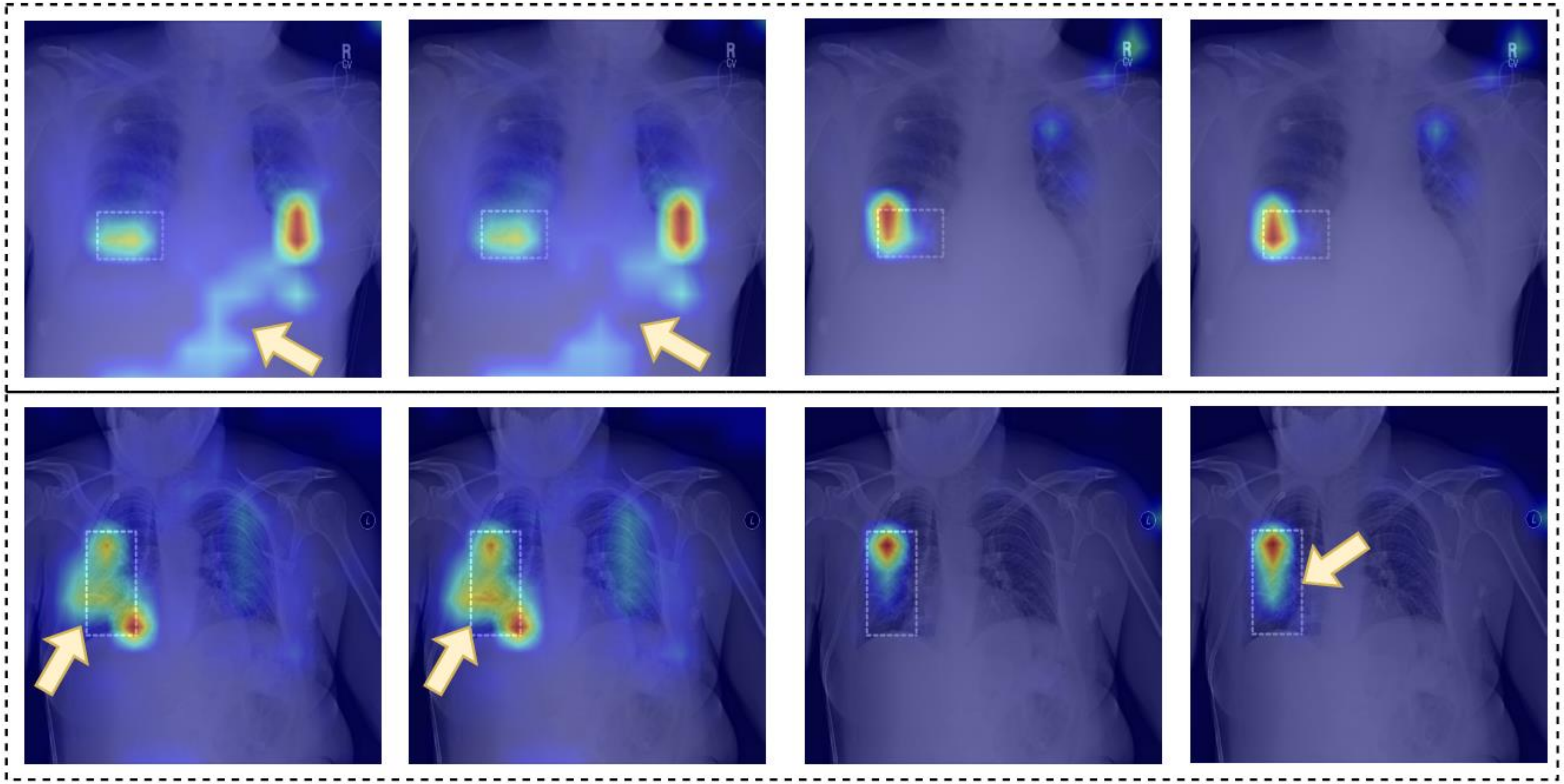
# Clinical Impact

- In US, approximately **530,000 general thoracic surgeries** are performed annually. [1]
- The prevalence of atelectasis accounts for **30.00–75.00%** of ordinary thoracic surgery [2]
- Assuming 530,000 operations performed, and 60% of atelectasis happened post-surgery.
- More than **15,900** patients with atelectasis are detected annually due to the increased sensitivity.
  - The sensitivity for H-GEL on Atelectasis is **0.697** compared to **0.649** for GAIN and 0.638 for KAD.

[1] Byrd, Catherine T., Kiah M. Williams, and Leah M. Backhus. "A brief overview of thoracic surgery in the United States." *Journal of Thoracic Disease* 14.1 (2022): 218.
[2] Zhao, Yongsheng, et al. "Systematic review and meta-analysis on perioperative intervention to prevent postoperative atelectasis complications after thoracic surgery." *Annals of Palliative Medicine* 10.10 (2021): 107260734-107210734.

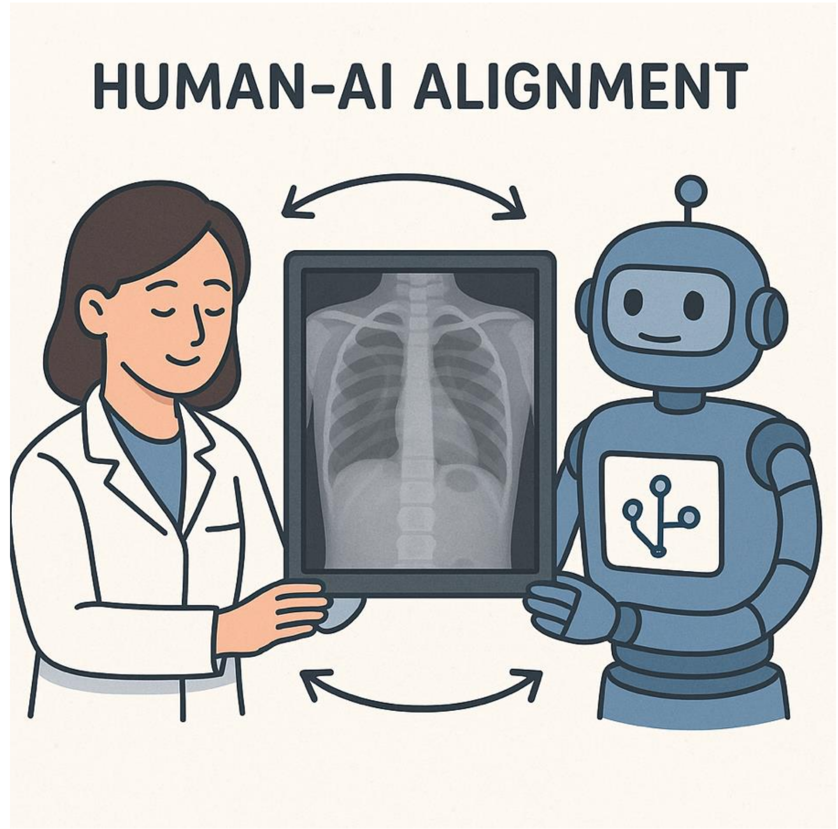
# H-EGL shows better alignment with human attention on the attention map.

- Model trained with DAL and H-EGL shows a clear reduction in false positive highlights
- Enhances **reliability** of the model's visual explanations



Baseline $\mathcal{L}_{\mathrm{HA}}$ $\mathcal{L}_{\mathrm{DAL}}$ $L_{\text{H-EGL}}$ ($\alpha$=1, $\beta$=1)

# Discussion and Further Work

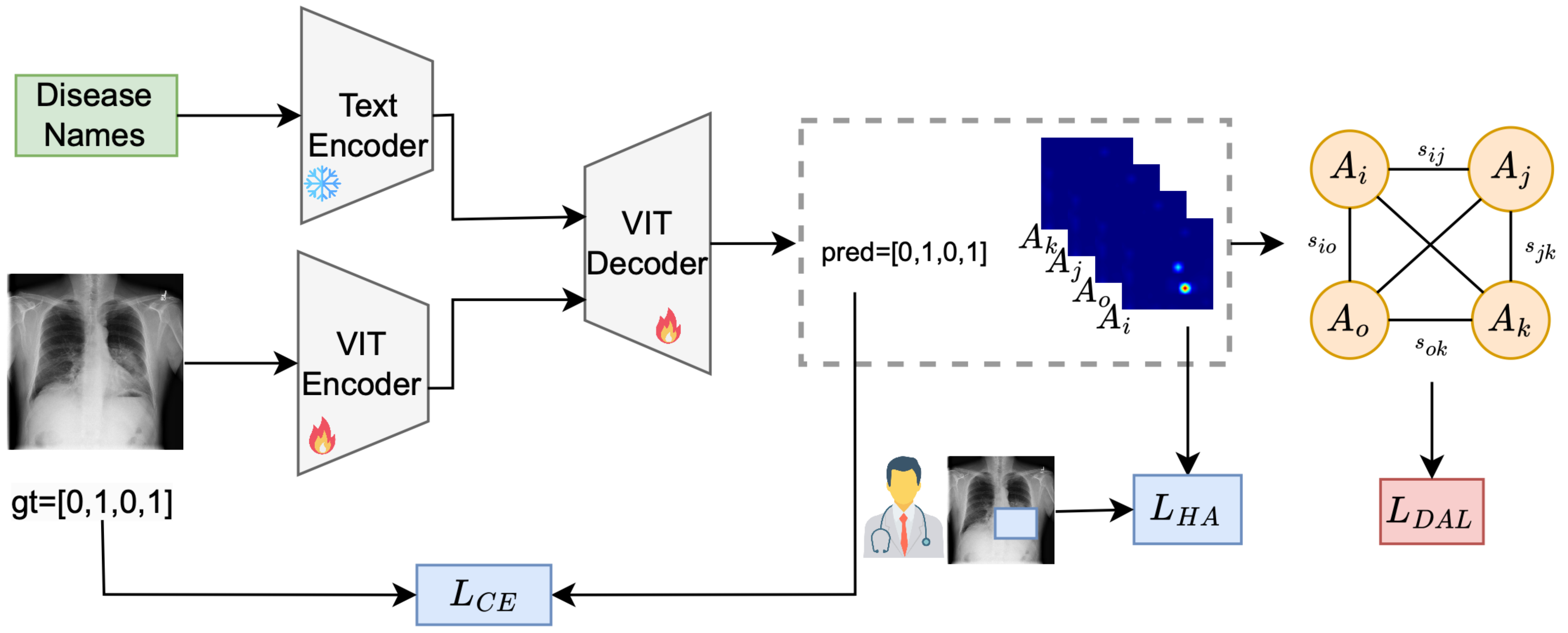



- H-EGL demonstrates strong potential for **combining self-supervised learning with human-guided attention alignment**
- This combination leads to improved accuracy and generalization
- Attention maps generated by H-EGL show **good interpretability and human alignment**, offering insights into the model's decision-making process
- DAL (self-supervised component) promotes **class-specific attention** without the need for localization labels, supporting **flexibility and scalability**
- Explore an optimal balance between **α and β** to further refine performance

# HYBRID EXPLANATION-GUIDED LEARNING FOR TRANSFORMER-BASED CHEST X-RAY DIAGNOSIS

**Shelley Zixin Shu**[1], Haozhe Luo[15], Alexander Poellinger[23], and Mauricio Reyes[14]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50, Bern 3008, Switzerland
[2] Inselspital (Bern University Hospital), 3010 Bern, Switzerland
[3] Insel Gruppe Bern Universitätsinstitut für Diagnostische, Interventionelle und Pädiatrische Radiologie
[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[5] Kaiko.AI, Zurich, Switzerland

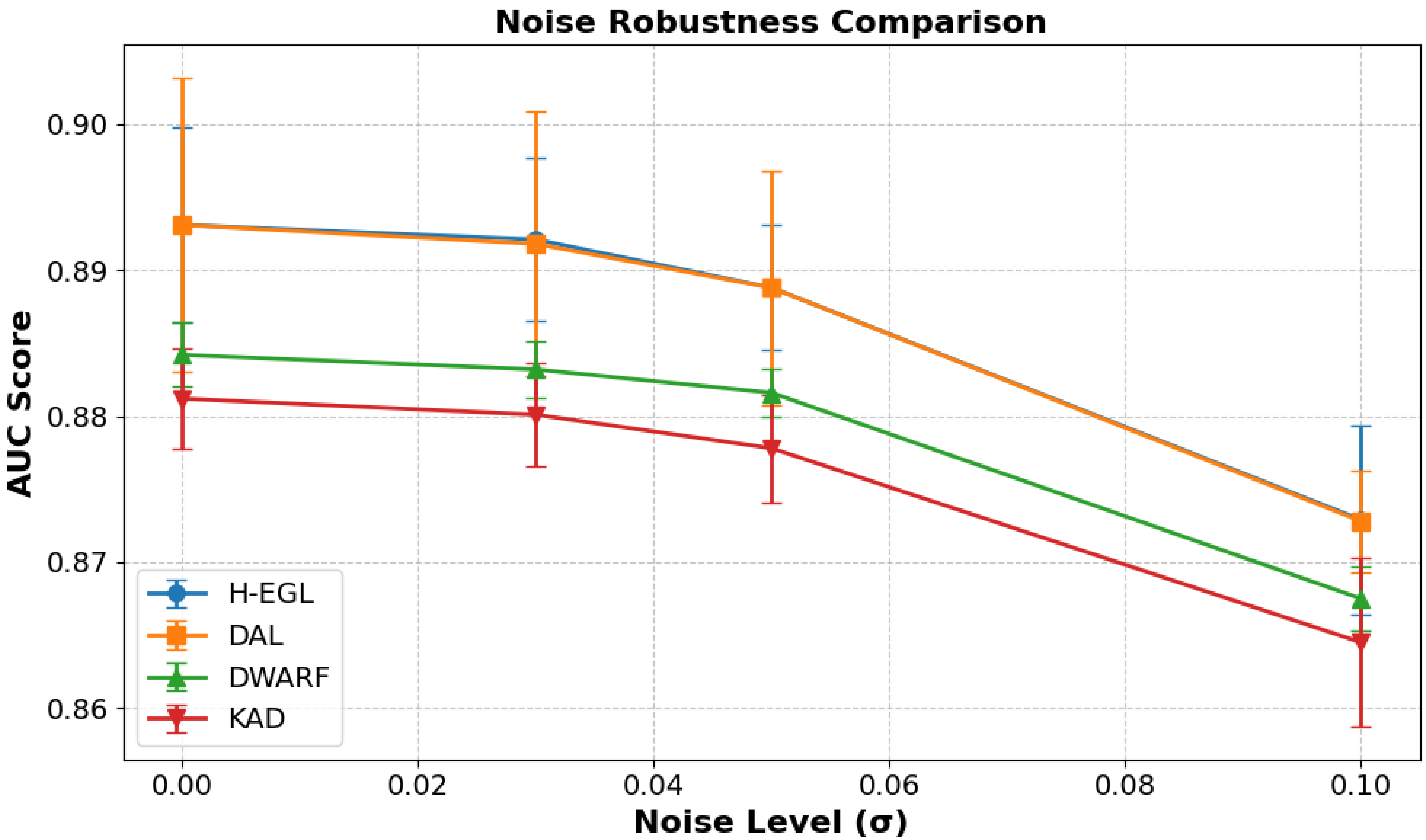


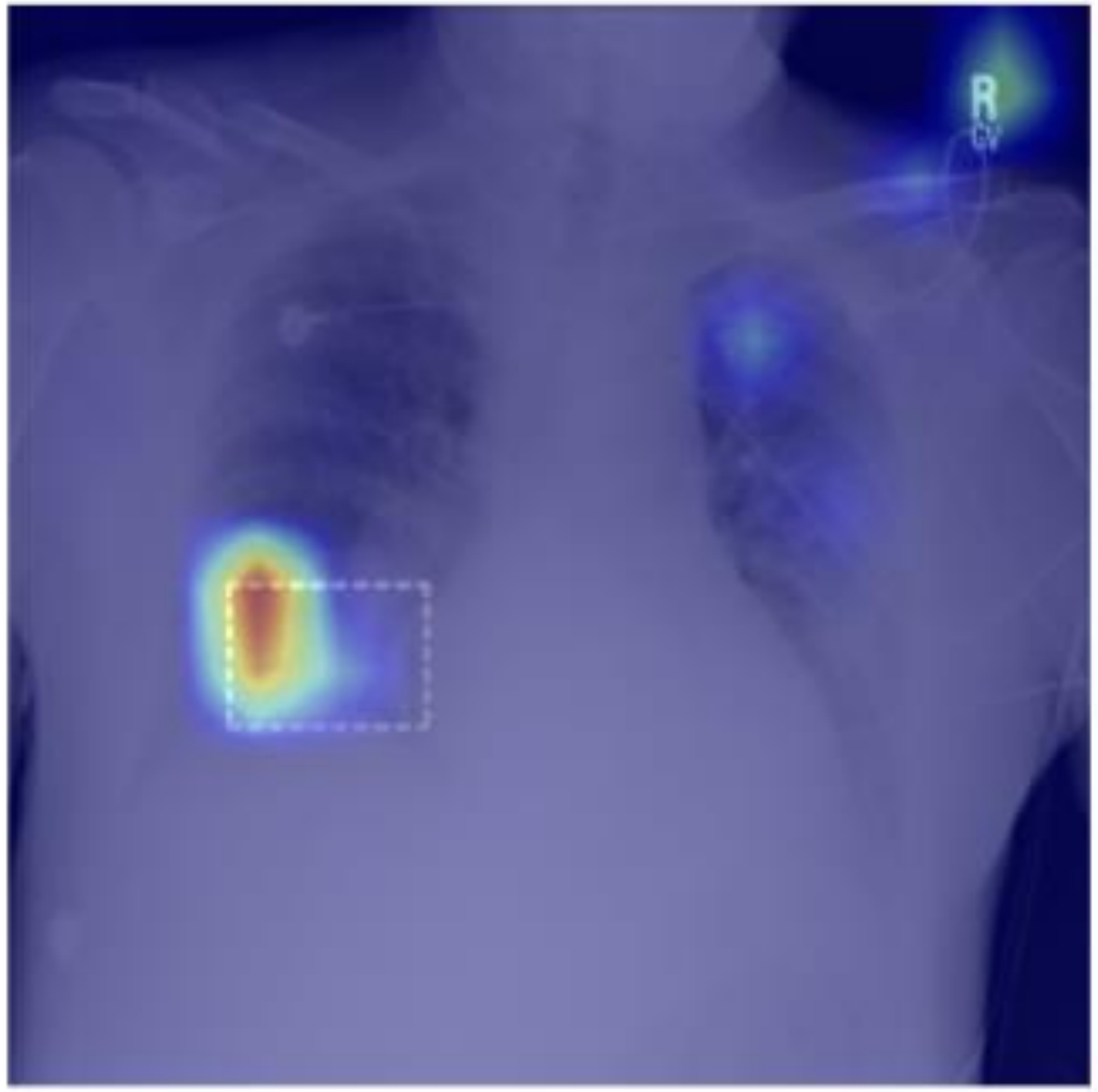

$$L_{\text{H-EGL}} = L_{CE} + \alpha L_{HA} + \beta L_{DAL}$$

Contact: Shelley Zixin Shu, zixin.shu@unibe.ch

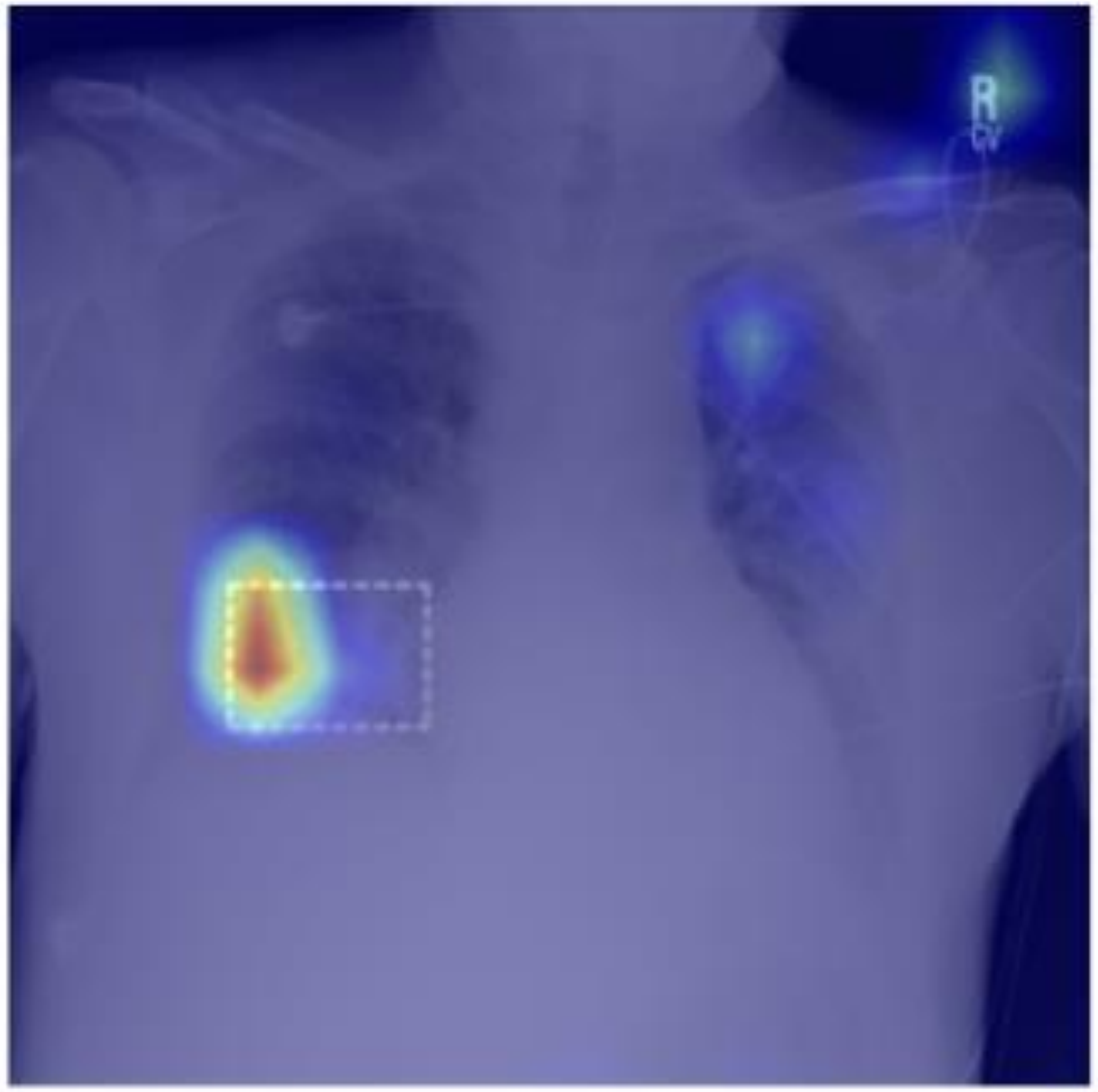
# H-EGL Demonstrates Strong Resilience to Noisy Inputs

— Adding normally distributed noise on the test image at inference time with various σ value.



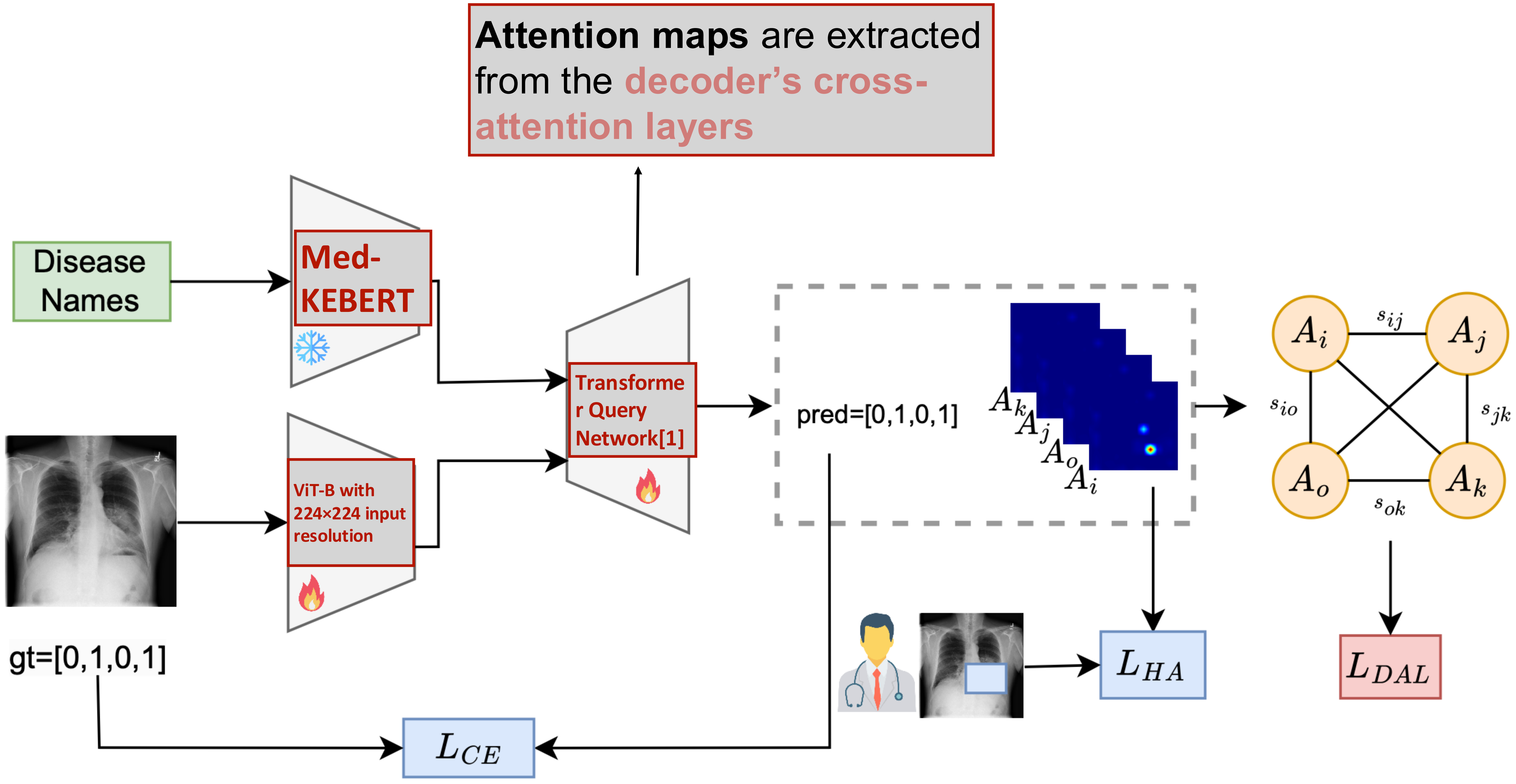


$\mathcal{L}_{\text{DAL}}$



$L_{\text{H-EGL}}$

# Implementation




$$L_{\text{H-EGL}} = L_{CE} + \alpha L_{HA} + \beta L_{DAL}$$

[1] Zhang, Xiaoman, et al. "Knowledge-enhanced visual-language pre-training on chest radiology images." *Nature Communications* 14.1 (2023): 4542.