ETH zürich

D-INFK

AI in Healthcare: A multimodal Perspective

Thomas Sutter
ETH Zurich

2025/06/05

medical
data
science

ETH zürich

D-INFK

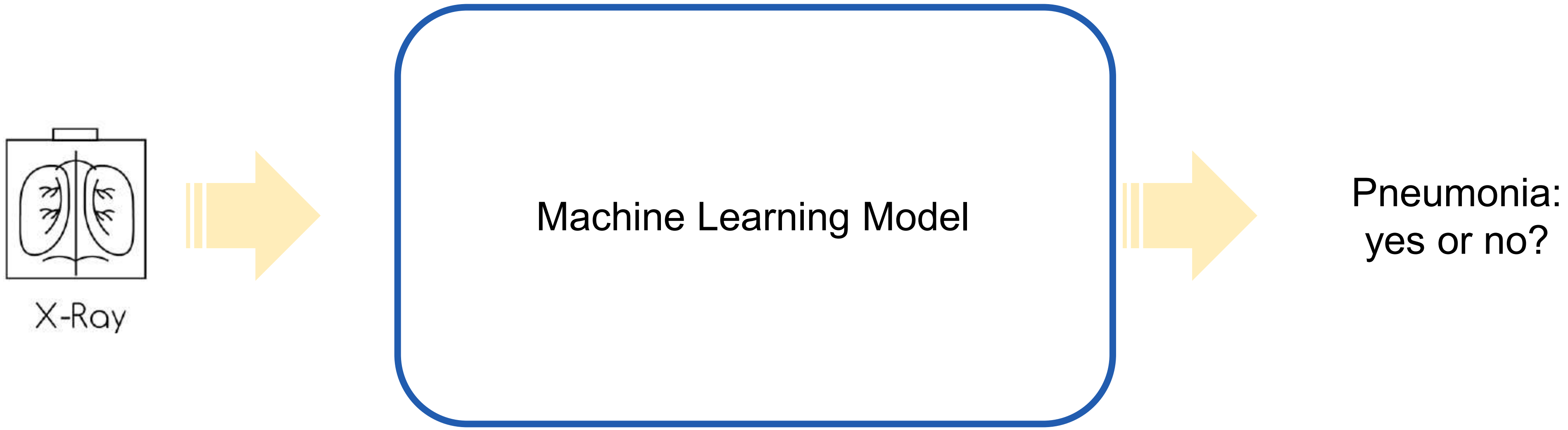
AI in Specialized Domains: A multimodal Perspective with applications in Healthcare

Thomas Sutter
ETH Zurich

2025/06/05

medical
data
science

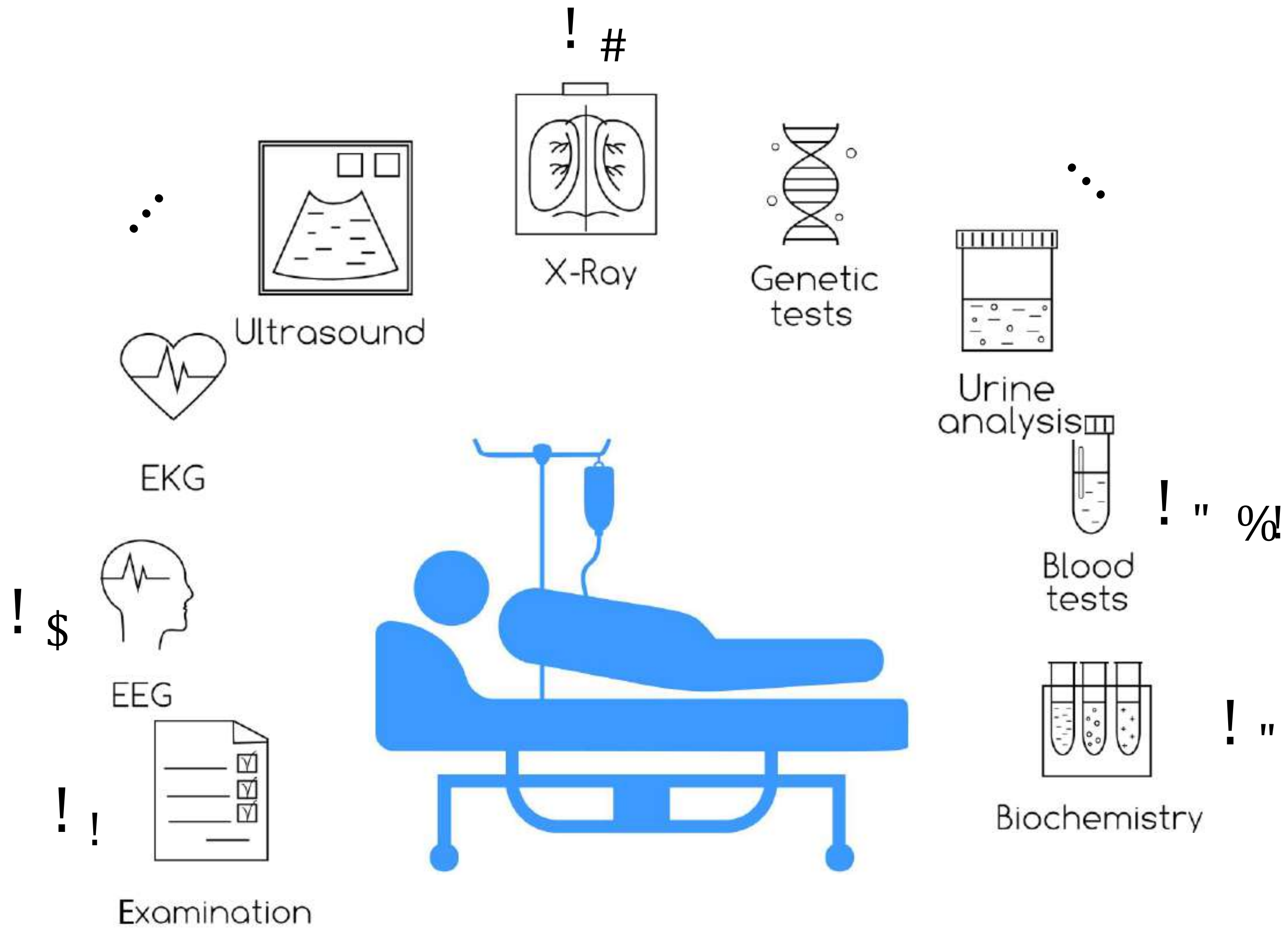
Machine Learning for Healthcare



Machine Learning Model

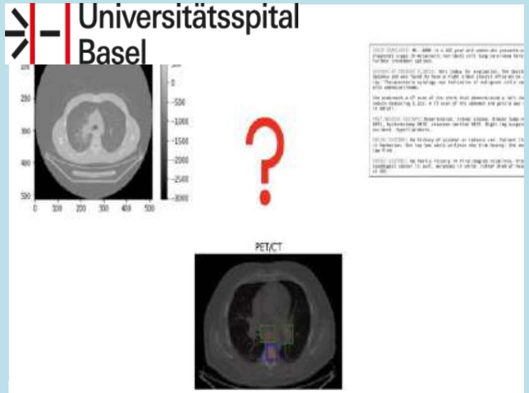
Pneumonia:
yes or no?

Healthcare: A multimodal perspective



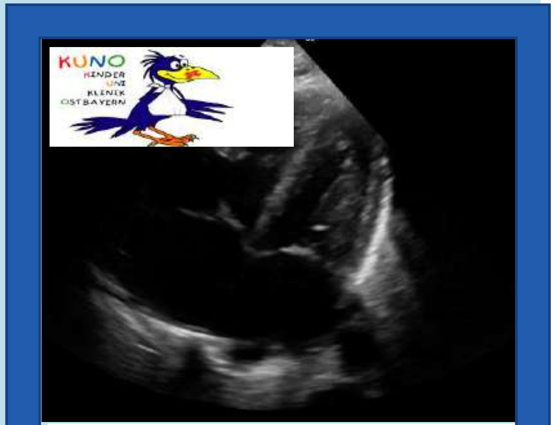
Examples of Multimodal Medical Applications

MIMIC- CXR

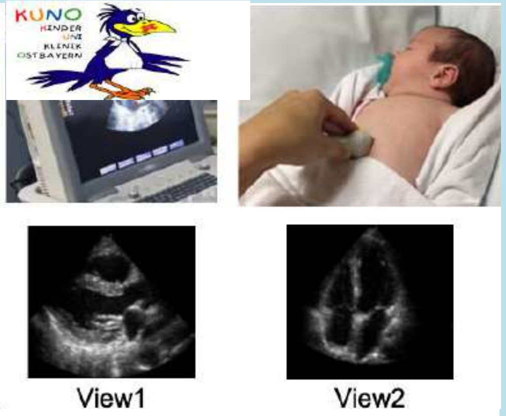


Lung Cancer

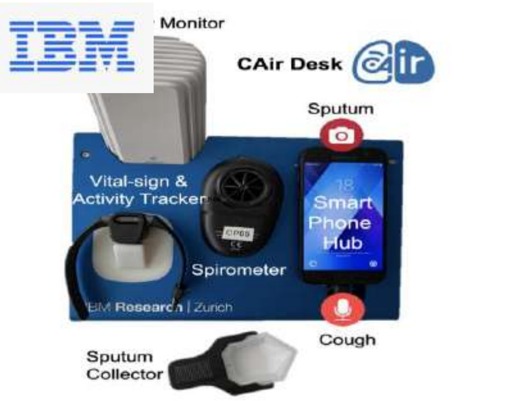
Appendicitis



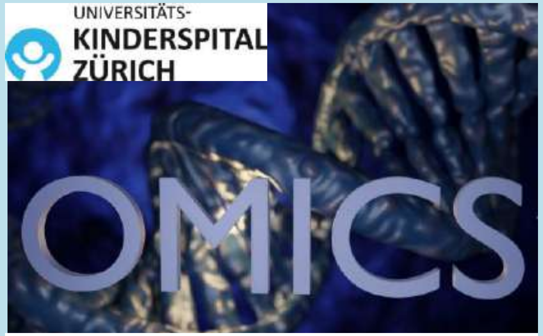
Heart-Defects



PH



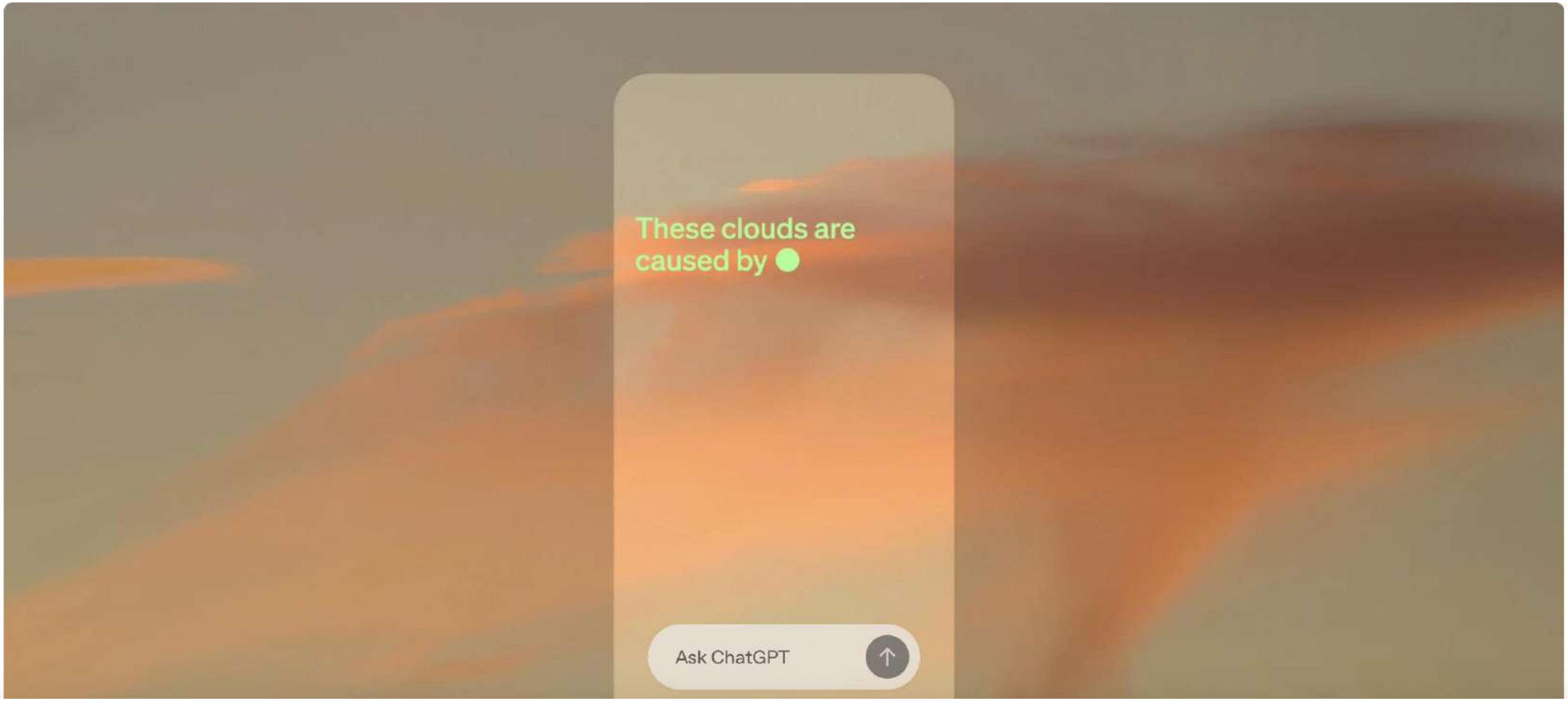
COPD



Multi-omics
Rare Diseases

Remote Home Monitoring

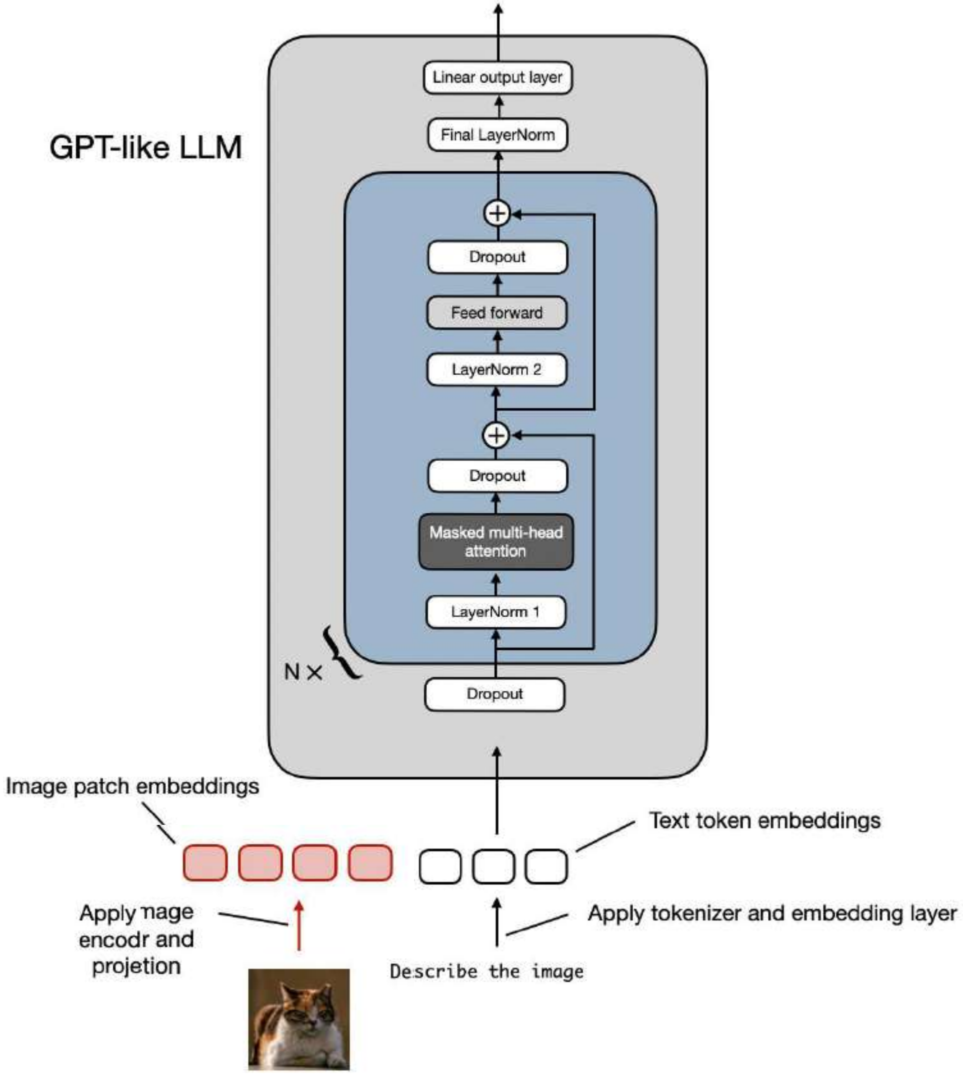
ChatGPT can now see, hear, and speak

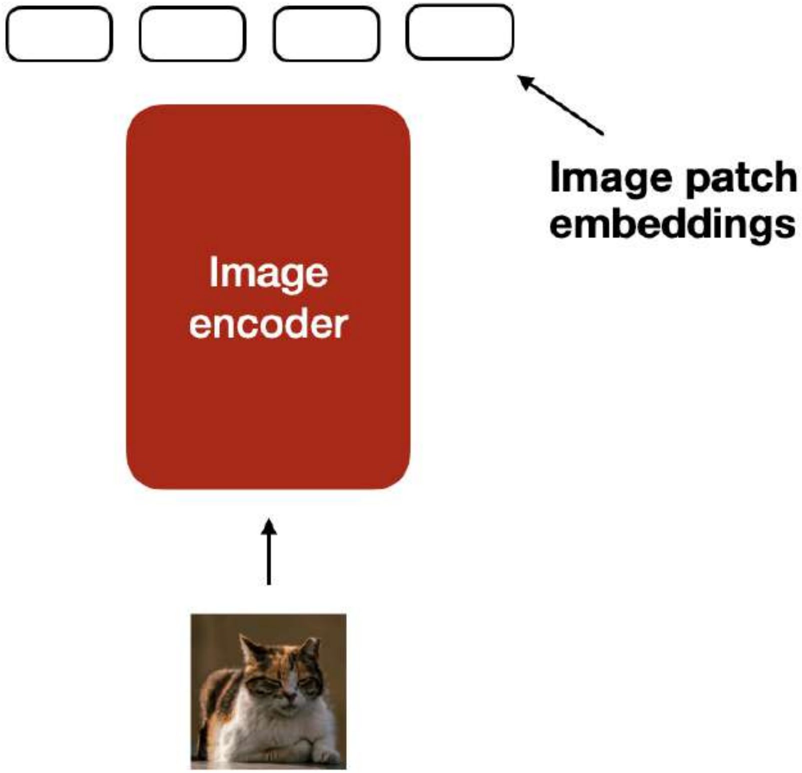


These clouds are caused by ●

https://openai.com/index/chatgpt-can-now-see-hear-and-speak/

Multimodality in modern AI Models



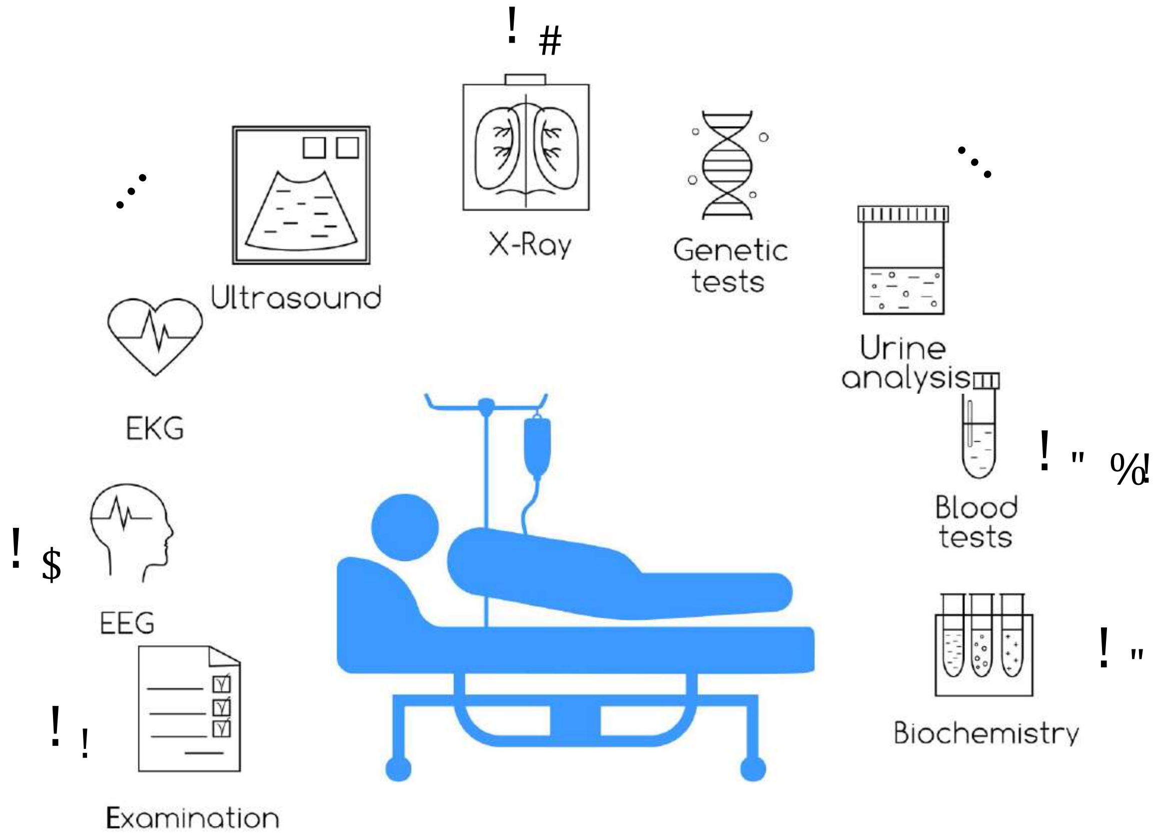


e.g., CLIP [1]

https://magazine.sebastianraschka.com/p/understanding-multimodal-llms

[1] Radford et al., «Learning Transferable Visual Models From Natural Language Supervision», ICML, 2021

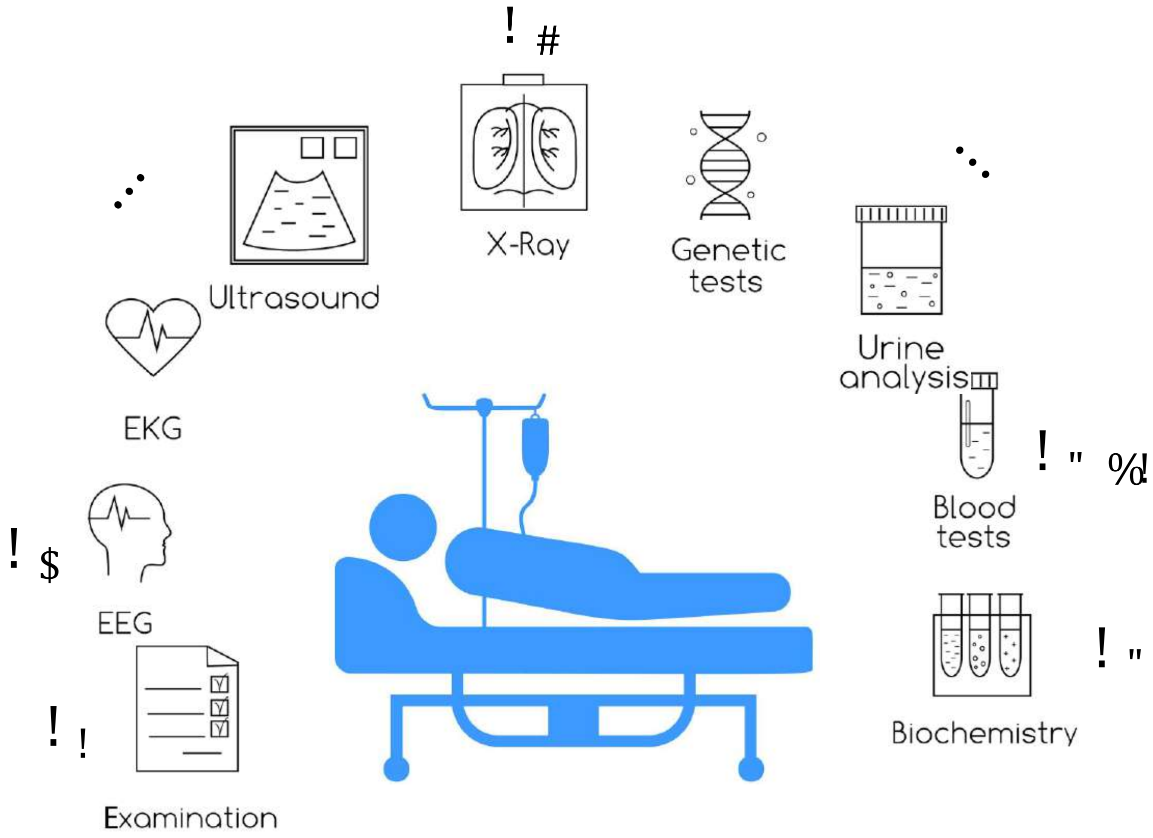
Healthcare Data



- ”small” scale
- Missingness
- Privacy Concerns
- Heterogeneity
- Expensive Annotation
- Challenging and different data types

Leveraging the structure of the data

Multimodal Learning under Weak Supervision



Weak Supervision
Learn from data without label annotation

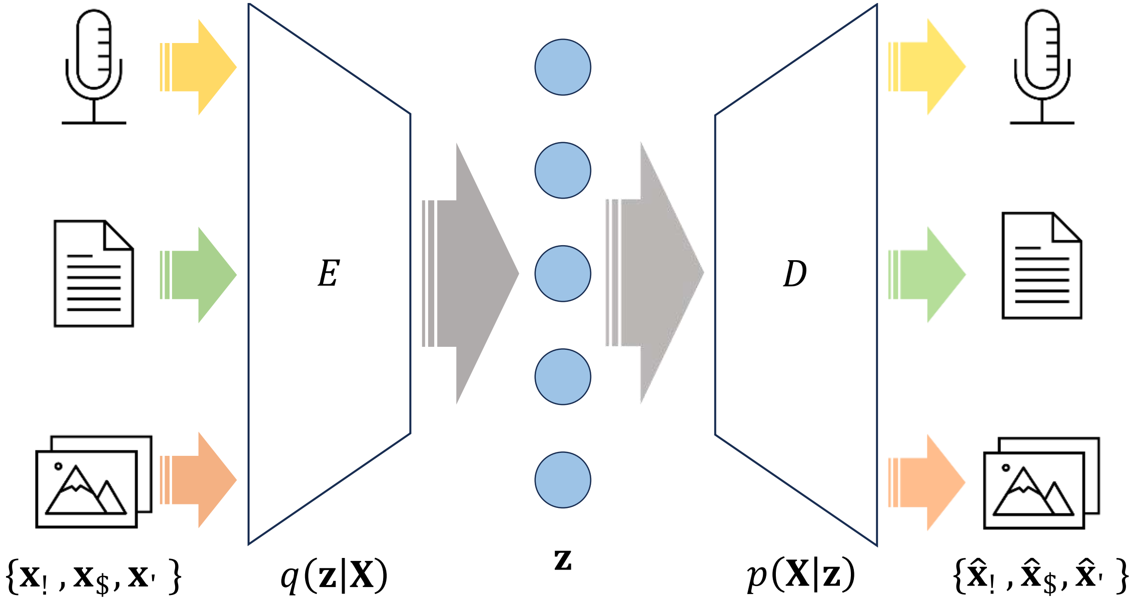
Goals

- Learn meaningful representations
- Be robust to missing modalities

1. **Sutter** et al, «Multimodal Learrning utilizing the Jensen-Shannon Divergence», Neurips 2020
2. Daunhawer, **Sutter**, Vogt, «Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models», DAGM GCPR, 2020
3. **Sutter** et al., «Generalized Multimodal ELBO», ICLR 2021
4. Klug, **Sutter**, Vogt, «Multimodal Generative Learning on the MIMIC-CXR Database», MIDL 2021
5. Daunhawer, **Sutter**, et al., «On the Limitations of Multimodal VAEs», ICLR 2021
6. **Sutter** et al., «Unity by Diversity: Improved Representation Learning for Multimodal VAEs», Neurips 2024
7. Agostini, …, Vogt and **Sutter**, «Weakly-Supervised Multimodal Learning on MIMIC-CXR», under submission, 2024

Multimodal Variational Autoencoders

- extension of the standard Variational Autoencoder [1]
- enables joint integration and reconstruction of two or more modalities


$\{\mathbf{x}_! , \mathbf{x}_\$, \mathbf{x}_' \}$
E
$q(\mathbf{z}|\mathbf{X})$
$\mathbf{z}$
D
$p(\mathbf{X}|\mathbf{z})$
$\{\hat{\mathbf{x}}_! , \hat{\mathbf{x}}_\$, \hat{\mathbf{x}}_' \}$

ELBO:

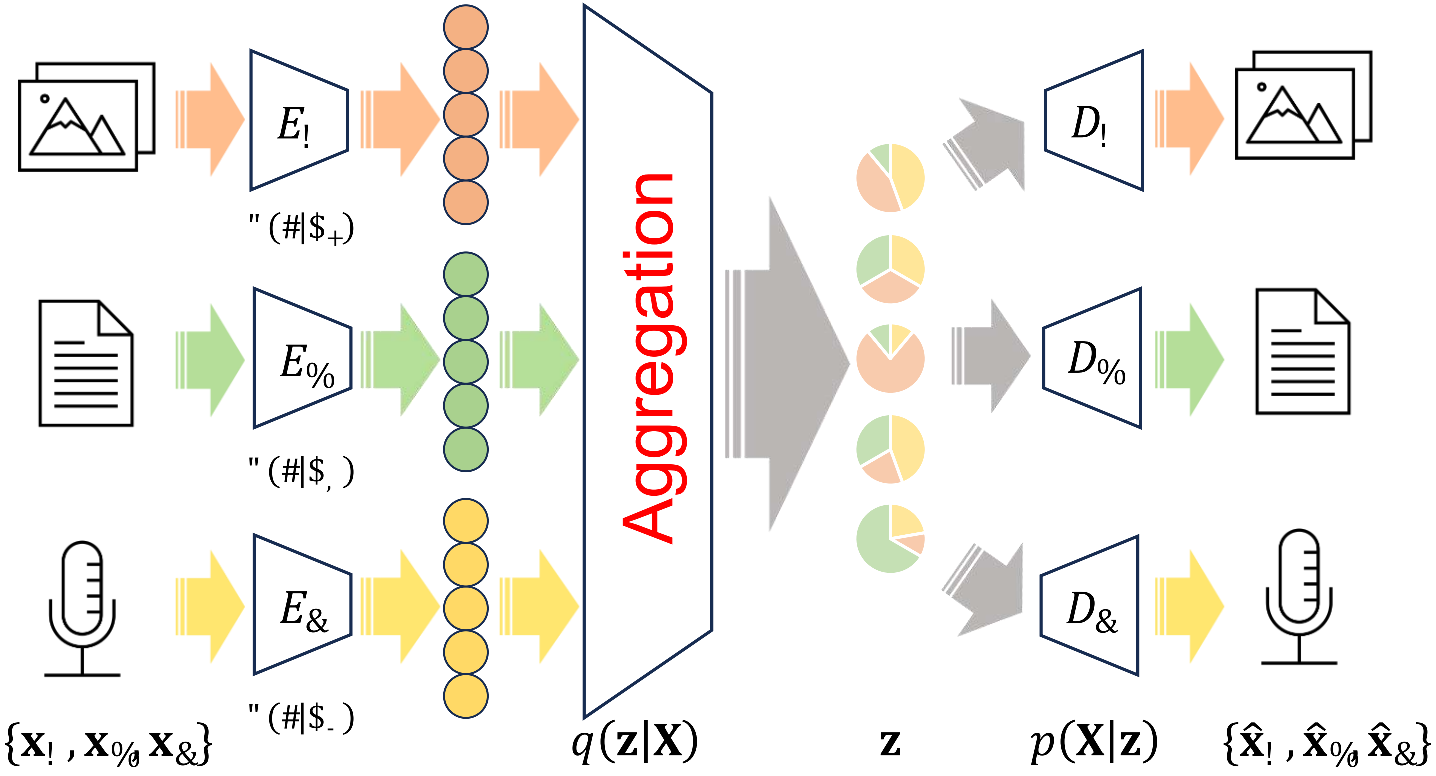
$$\log p(\boldsymbol{X}) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{X})}\left[\log p(\boldsymbol{X} \mid \boldsymbol{z}) - \log \frac{q(\boldsymbol{z} \mid \boldsymbol{X})}{p(\boldsymbol{z})}\right]$$

- $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$: multimodal sample
- $\boldsymbol{x}_m$: sample of modality m
- $p(\boldsymbol{X} \mid \boldsymbol{z})$: probability of a sample $\boldsymbol{X}$ given the latent vector $\boldsymbol{z}$
- $q(\boldsymbol{z} \mid \boldsymbol{X})$: posterior approximation of $\boldsymbol{z}$
- $p(\boldsymbol{z})$: prior distribution of $\boldsymbol{z}$

[1] Kingma, Welling, Auto-Encoding Variational Bayes, ICLR, 2014

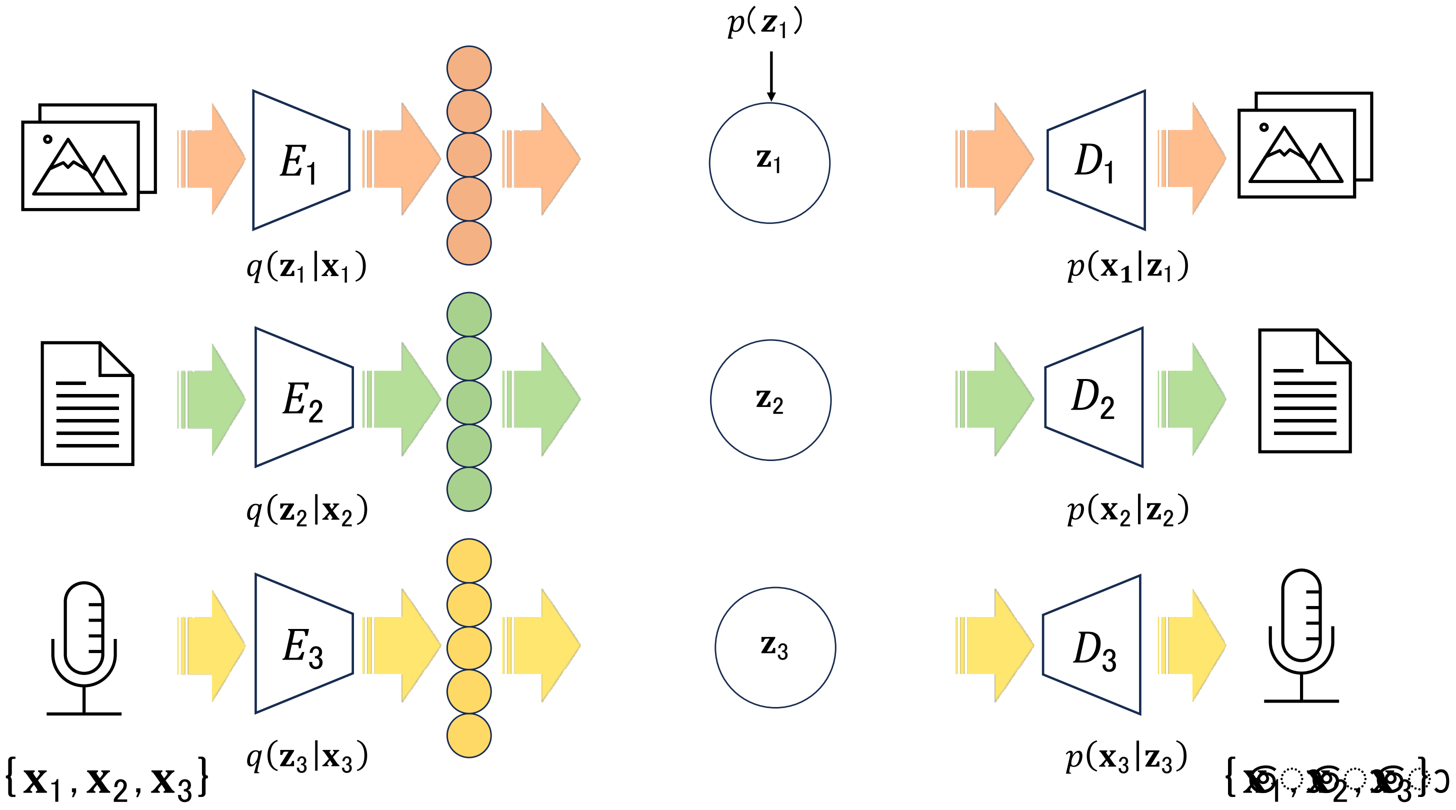
Learning a Joint Multimodal Representation

$$\boxed{\mathcal{E}(\boldsymbol{X}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log \frac{q_\phi(\boldsymbol{z} \mid \boldsymbol{X})}{p_\theta(\boldsymbol{z})}\right]}$$



Set of Independent VAEs

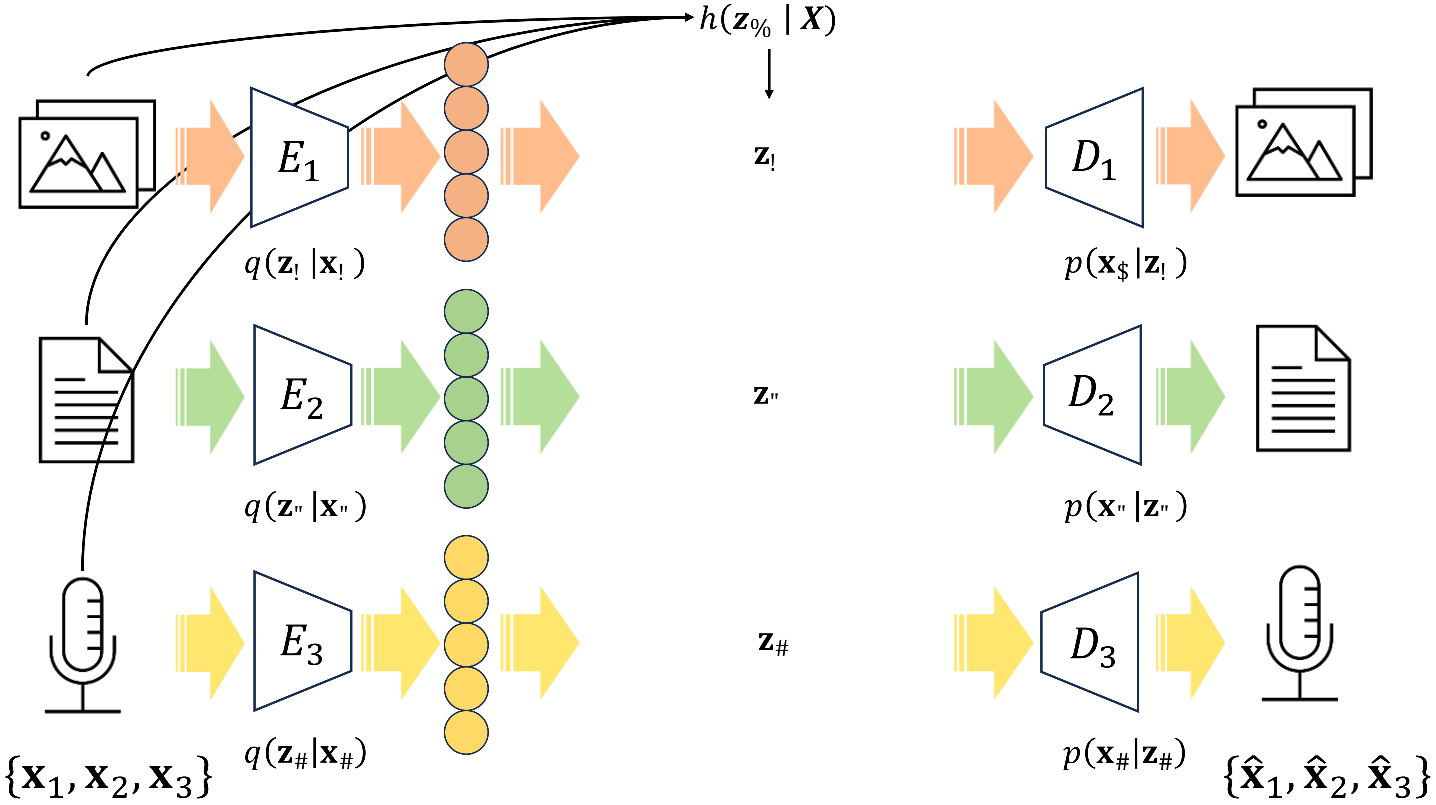
$$\mathcal{E}(\boldsymbol{X}) = \sum_{m=1}^{M} \mathbb{E}_{q^m(\boldsymbol{z}_m|\boldsymbol{x}_m)}\left[\log p(\boldsymbol{x}_m \mid \boldsymbol{z}_m) - \log \frac{q^m(\boldsymbol{z}_m \mid \boldsymbol{x}_m)}{p(\boldsymbol{z}_m)}\right]$$



$p(\mathbf{z}_1)$
E_1
$\mathbf{z}_1$
D_1
$q(\mathbf{z}_1|\mathbf{x}_1)$
$p(\mathbf{x}_1|\mathbf{z}_1)$
E_2
$\mathbf{z}_2$
D_2
$q(\mathbf{z}_2|\mathbf{x}_2)$
$p(\mathbf{x}_2|\mathbf{z}_2)$
E_3
$\mathbf{z}_3$
D_3
$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$
$q(\mathbf{z}_3|\mathbf{x}_3)$
$p(\mathbf{x}_3|\mathbf{z}_3)$
$\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{x}}_3\}$

Multimodal Variational Mixture Prior (MMVM)

$$\mathcal{E}(\boldsymbol{X}) = \sum_{m=1}^{M} \mathbb{E}_{q^m(\boldsymbol{z}_m|\boldsymbol{x}_m)}\left[\log p(\boldsymbol{x}_m \mid \boldsymbol{z}_m) - \log \frac{q^m(\boldsymbol{z}_m \mid \boldsymbol{x}_m)}{\boxed{h(\boldsymbol{z}_m \mid \boldsymbol{X})}}\right]$$



$h(\mathbf{z}_{\%} \mid \boldsymbol{X})$
E_1
E_2
E_3
$q(\mathbf{z}_! | \mathbf{x}_!)$
$q(\mathbf{z}_" | \mathbf{x}_")$
$q(\mathbf{z}_\# | \mathbf{x}_\#)$
$\mathbf{z}_!$
$\mathbf{z}_"$
$\mathbf{z}_\#$
D_1
D_2
D_3
$p(\mathbf{x}_\$ | \mathbf{z}_!)$
$p(\mathbf{x}_" | \mathbf{z}_")$
$p(\mathbf{x}_\# | \mathbf{z}_\#)$
$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$
$\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{x}}_3\}$

MMVM VAE

From a sum of unimodal ELBOs to the MMVM-prior objective

$$\mathcal{E}(\boldsymbol{X}) = \sum_{m=1}^{M} \mathbb{E}_{q^m(\boldsymbol{z}_m \mid \boldsymbol{x}_m)} \left[\log p(\boldsymbol{x}_m \mid \boldsymbol{z}_m) - \log \frac{q^m(\boldsymbol{z}_m \mid \boldsymbol{x}_m)}{\frac{1}{M} \sum_{\tilde{m}=1}^{M} q^{\tilde{m}}(\boldsymbol{z}_m \mid \boldsymbol{x}_{\tilde{m}})} \right]$$

We introduce the MMVM prior distributions [1]

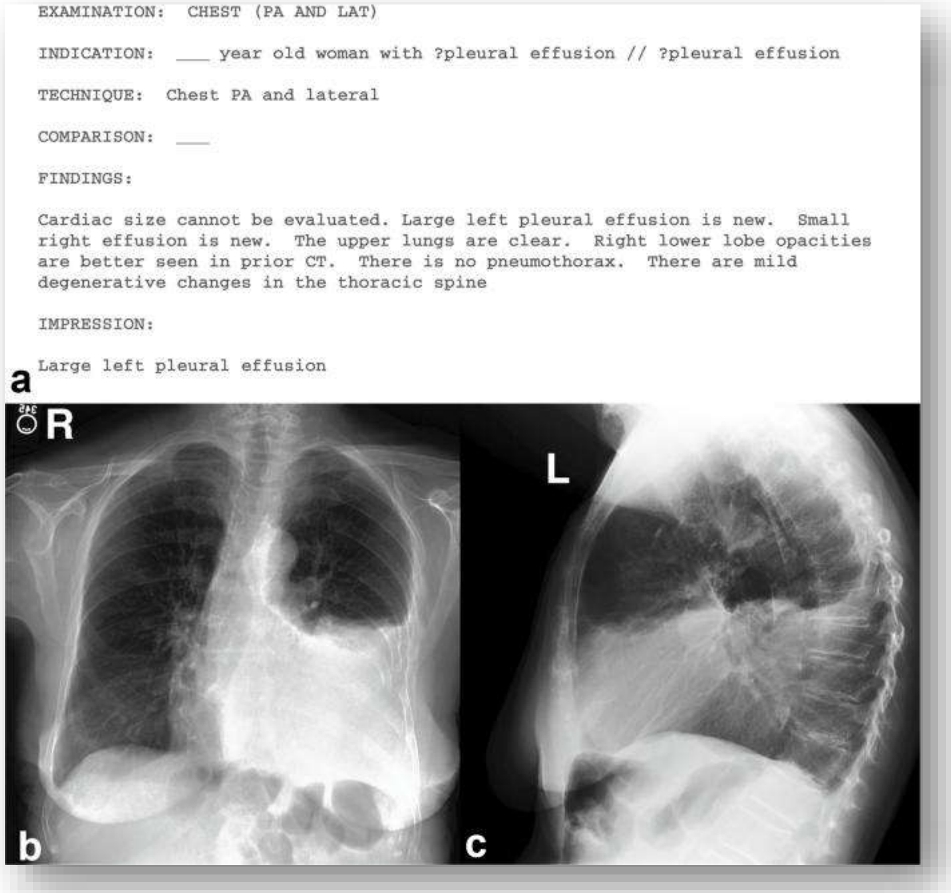
$$p(\boldsymbol{z}_m) = h(\boldsymbol{z}_m \mid \boldsymbol{X}) = \frac{1}{M} \sum_{\tilde{m}=1}^{M} q^{\tilde{m}}(\boldsymbol{z}_m \mid \boldsymbol{x}_{\tilde{m}})$$

Inspired by the VAMP prior [2], we can show optimality of the chosen prior distribution.

[1] Sutter et al., Unity by Diversity: Improved Representation Learning in Multimodal VAEs, Neurips 2024
[2] Tomczak and Welling, VAE with a VAMP prior, AISTATS 2018

Mimic-CXR

MIMIC-CXR



EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion

- MIMIC-CXR is a large **publicly available dataset of chest radiographs**[1]

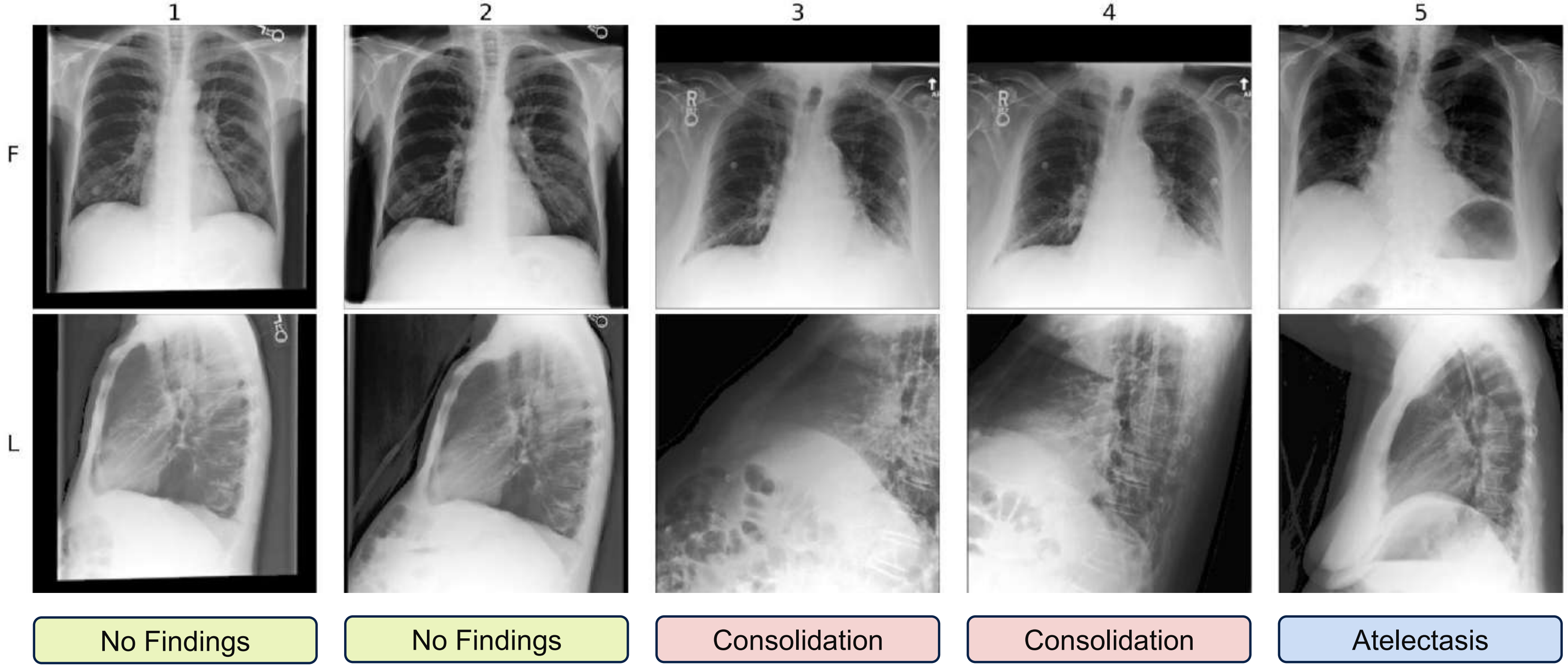
- A total of **377.110 images** corresponding to **227.835 studies**

- **Multimodal**:
  - ❖ Images from multiple view positions
  - ❖ Radiology reports in text form
  - ❖ Electronic Health Records

[1] Johnson et al., «MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports», Sci Data, 2019
[2] Agostini, …, Vogt and **Sutter**, «Weakly-Supervised Multimodal Learning on MIMIC-CXR», ML4H, 2024

Bimodal Mimic-CXR Dataset

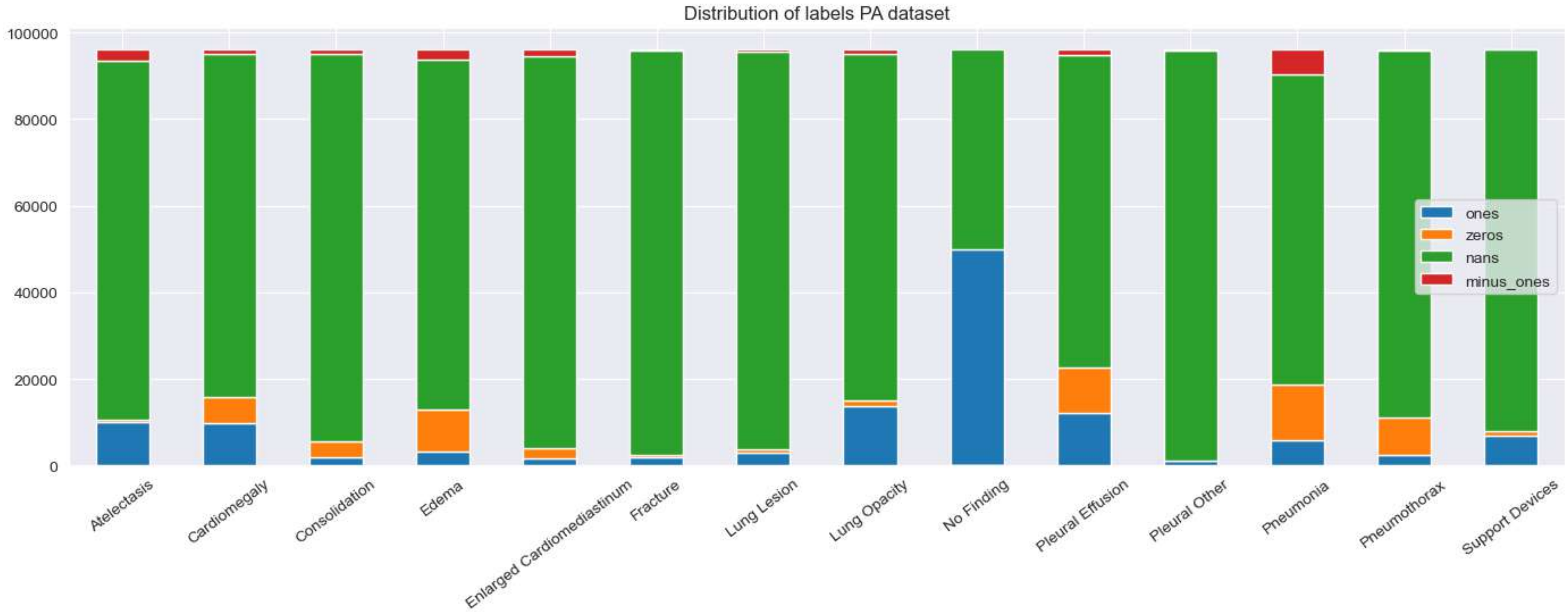
- $\boldsymbol{F}: \{'PA', 'AP'\}, \ \boldsymbol{L}: \{'Lat', 'LL'\}$
- $Dataset: \mathbf{X} = \{\mathrm{X}^{(\mathrm{i})}\}_{\mathrm{i}=1}^{\mathrm{n}}, \ \mathrm{X}^{(\mathrm{i})} = \{\mathrm{x}_{f}^{(\mathrm{i})}, \mathrm{x}_{l}^{(\mathrm{i})}\}$



[1] Johnson et al., «MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports», Sci Data, 2019
[2] Agostini, …, Vogt and **Sutter**, «Weakly-Supervised Multimodal Learning on MIMIC-CXR», ML4H, 2024

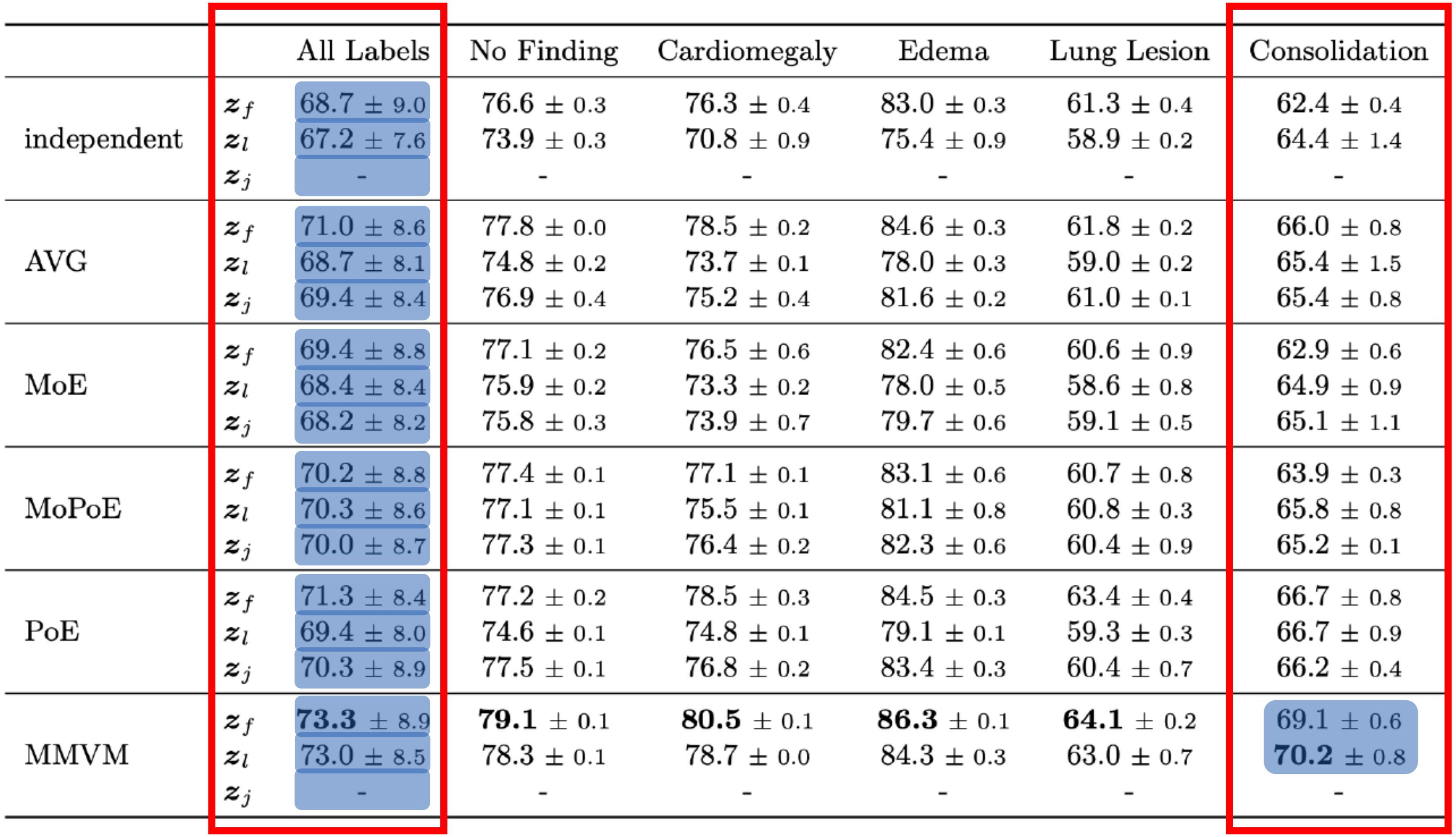
MIMIC-CXR Labels

- Multiclass Labels are generated from radiology reports - **14 diseases** and **4 classes**
- Labels are usually **binarized** [1, 2]



[1] Seyyed-Kalantari, Laleh, et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers." *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. 2020.
[2] Haque, Md Inzamam Ul, et al. "Effect of image resolution on automated classification of chest X-rays." *Journal of Medical Imaging* 10.4 (2023): 044503-044503.

MIMIC-CXR: Comparison with other VAEs

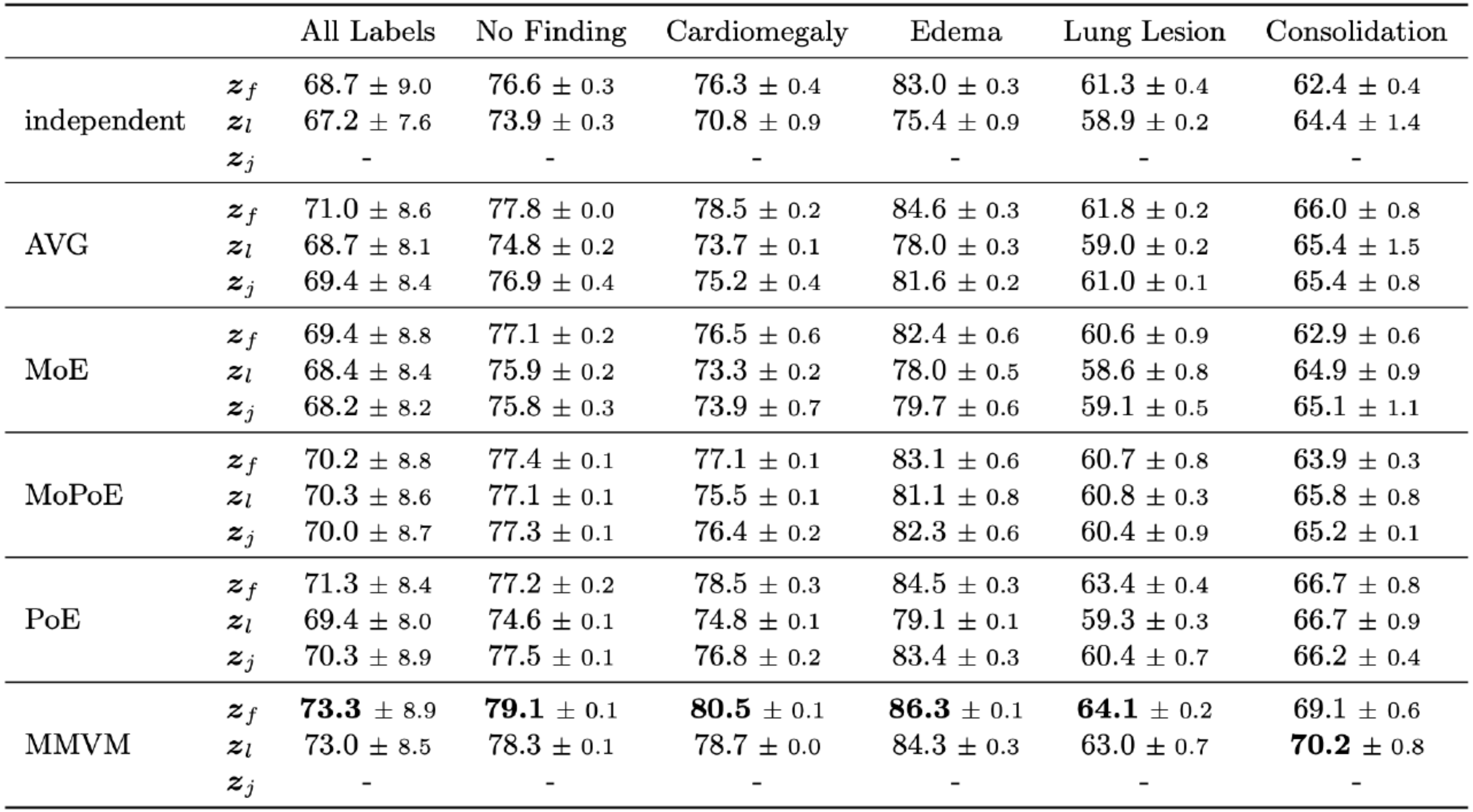


		All Labels	No Finding	Cardiomegaly	Edema	Lung Lesion	Consolidation
independent	z_f	68.7 ± 9.0	76.6 ± 0.3	76.3 ± 0.4	83.0 ± 0.3	61.3 ± 0.4	62.4 ± 0.4
	z_l	67.2 ± 7.6	73.9 ± 0.3	70.8 ± 0.9	75.4 ± 0.9	58.9 ± 0.2	64.4 ± 1.4
	z_j	-	-	-	-	-	-
AVG	z_f	71.0 ± 8.6	77.8 ± 0.0	78.5 ± 0.2	84.6 ± 0.3	61.8 ± 0.2	66.0 ± 0.8
	z_l	68.7 ± 8.1	74.8 ± 0.2	73.7 ± 0.1	78.0 ± 0.3	59.0 ± 0.2	65.4 ± 1.5
	z_j	69.4 ± 8.4	76.9 ± 0.4	75.2 ± 0.4	81.6 ± 0.2	61.0 ± 0.1	65.4 ± 0.8
MoE	z_f	69.4 ± 8.8	77.1 ± 0.2	76.5 ± 0.6	82.4 ± 0.6	60.6 ± 0.9	62.9 ± 0.6
	z_l	68.4 ± 8.4	75.9 ± 0.2	73.3 ± 0.2	78.0 ± 0.5	58.6 ± 0.8	64.9 ± 0.9
	z_j	68.2 ± 8.2	75.8 ± 0.3	73.9 ± 0.7	79.7 ± 0.6	59.1 ± 0.5	65.1 ± 1.1
MoPoE	z_f	70.2 ± 8.8	77.4 ± 0.1	77.1 ± 0.1	83.1 ± 0.6	60.7 ± 0.8	63.9 ± 0.3
	z_l	70.3 ± 8.6	77.1 ± 0.1	75.5 ± 0.1	81.1 ± 0.8	60.8 ± 0.3	65.8 ± 0.8
	z_j	70.0 ± 8.7	77.3 ± 0.1	76.4 ± 0.2	82.3 ± 0.6	60.4 ± 0.9	65.2 ± 0.1
PoE	z_f	71.3 ± 8.4	77.2 ± 0.2	78.5 ± 0.3	84.5 ± 0.3	63.4 ± 0.4	66.7 ± 0.8
	z_l	69.4 ± 8.0	74.6 ± 0.1	74.8 ± 0.1	79.1 ± 0.1	59.3 ± 0.3	66.7 ± 0.9
	z_j	70.3 ± 8.9	77.5 ± 0.1	76.8 ± 0.2	83.4 ± 0.3	60.4 ± 0.7	66.2 ± 0.4
MMVM	z_f	**73.3** ± 8.9	**79.1** ± 0.1	**80.5** ± 0.1	**86.3** ± 0.1	**64.1** ± 0.2	69.1 ± 0.6
	z_l	73.0 ± 8.5	78.3 ± 0.1	78.7 ± 0.0	84.3 ± 0.3	63.0 ± 0.7	**70.2** ± 0.8
	z_j	-	-	-	-	-	-

We report the AUROC of binary classification tasks.

[1] **Sutter** et al., «Unity by Diversity: Improved Representation Learning for Multimodal VAEs», Neurips 2024
[2] Agostini, …, Vogt and **Sutter**, «Weakly-Supervised Multimodal Learning on MIMIC-CXR», ML4H, 2024

MIMIC-CXR: Comparison with other VAEs

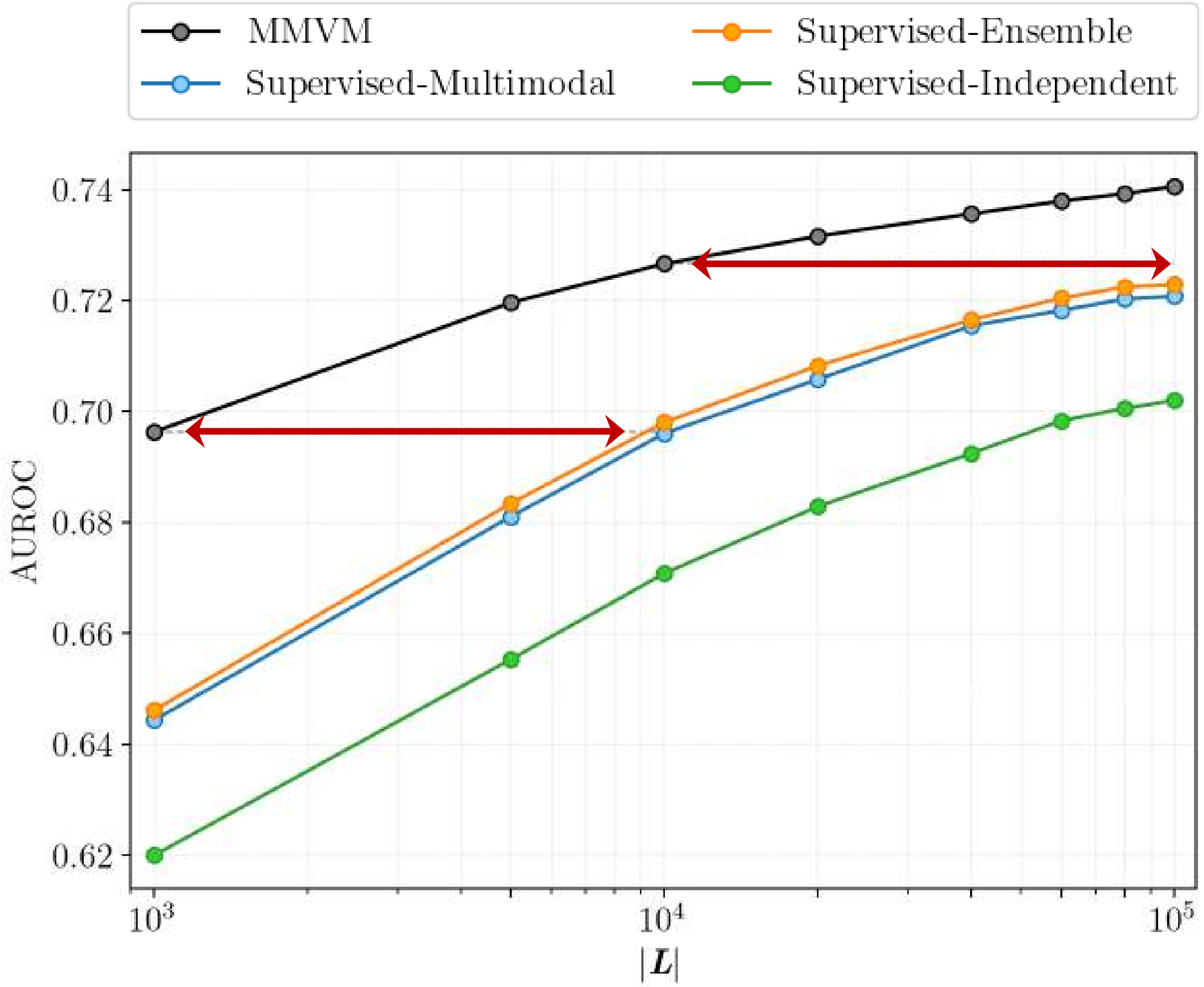


		All Labels	No Finding	Cardiomegaly	Edema	Lung Lesion	Consolidation
independent	$\boldsymbol{z}_f$	68.7 ± 9.0	76.6 ± 0.3	76.3 ± 0.4	83.0 ± 0.3	61.3 ± 0.4	62.4 ± 0.4
	$\boldsymbol{z}_l$	67.2 ± 7.6	73.9 ± 0.3	70.8 ± 0.9	75.4 ± 0.9	58.9 ± 0.2	64.4 ± 1.4
	$\boldsymbol{z}_j$	-	-	-	-	-	-
AVG	$\boldsymbol{z}_f$	71.0 ± 8.6	77.8 ± 0.0	78.5 ± 0.2	84.6 ± 0.3	61.8 ± 0.2	66.0 ± 0.8
	$\boldsymbol{z}_l$	68.7 ± 8.1	74.8 ± 0.2	73.7 ± 0.1	78.0 ± 0.3	59.0 ± 0.2	65.4 ± 1.5
	$\boldsymbol{z}_j$	69.4 ± 8.4	76.9 ± 0.4	75.2 ± 0.4	81.6 ± 0.2	61.0 ± 0.1	65.4 ± 0.8
MoE	$\boldsymbol{z}_f$	69.4 ± 8.8	77.1 ± 0.2	76.5 ± 0.6	82.4 ± 0.6	60.6 ± 0.9	62.9 ± 0.6
	$\boldsymbol{z}_l$	68.4 ± 8.4	75.9 ± 0.2	73.3 ± 0.2	78.0 ± 0.5	58.6 ± 0.8	64.9 ± 0.9
	$\boldsymbol{z}_j$	68.2 ± 8.2	75.8 ± 0.3	73.9 ± 0.7	79.7 ± 0.6	59.1 ± 0.5	65.1 ± 1.1
MoPoE	$\boldsymbol{z}_f$	70.2 ± 8.8	77.4 ± 0.1	77.1 ± 0.1	83.1 ± 0.6	60.7 ± 0.8	63.9 ± 0.3
	$\boldsymbol{z}_l$	70.3 ± 8.6	77.1 ± 0.1	75.5 ± 0.1	81.1 ± 0.8	60.8 ± 0.3	65.8 ± 0.8
	$\boldsymbol{z}_j$	70.0 ± 8.7	77.3 ± 0.1	76.4 ± 0.2	82.3 ± 0.6	60.4 ± 0.9	65.2 ± 0.1
PoE	$\boldsymbol{z}_f$	71.3 ± 8.4	77.2 ± 0.2	78.5 ± 0.3	84.5 ± 0.3	63.4 ± 0.4	66.7 ± 0.8
	$\boldsymbol{z}_l$	69.4 ± 8.0	74.6 ± 0.1	74.8 ± 0.1	79.1 ± 0.1	59.3 ± 0.3	66.7 ± 0.9
	$\boldsymbol{z}_j$	70.3 ± 8.9	77.5 ± 0.1	76.8 ± 0.2	83.4 ± 0.3	60.4 ± 0.7	66.2 ± 0.4
MMVM	$\boldsymbol{z}_f$	**73.3** ± 8.9	**79.1** ± 0.1	**80.5** ± 0.1	**86.3** ± 0.1	**64.1** ± 0.2	69.1 ± 0.6
	$\boldsymbol{z}_l$	73.0 ± 8.5	78.3 ± 0.1	78.7 ± 0.0	84.3 ± 0.3	63.0 ± 0.7	**70.2** ± 0.8
	$\boldsymbol{z}_j$	-	-	-	-	-	-

We report the AUROC of binary classification tasks.

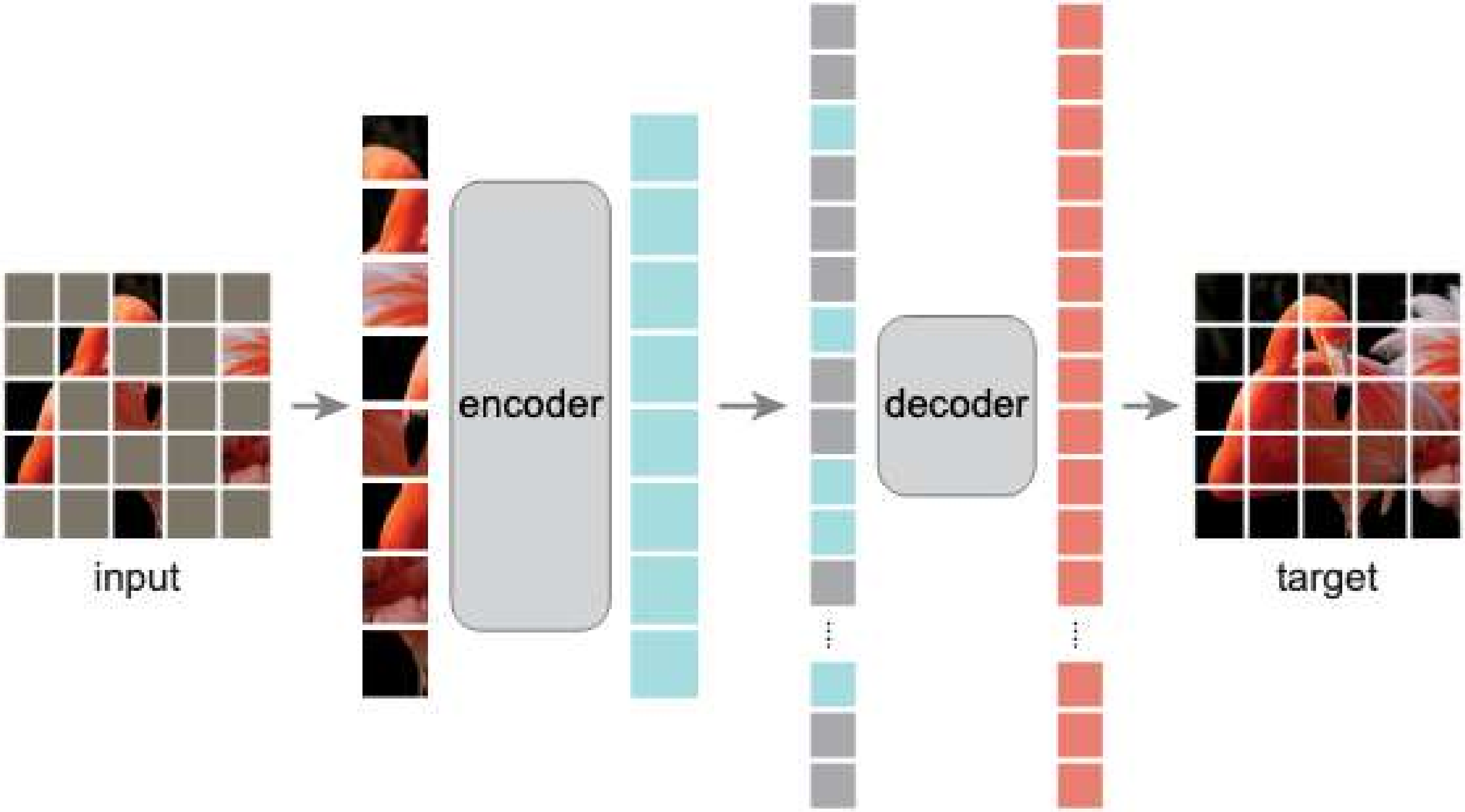
[1] **Sutter** et al., «Unity by Diversity: Improved Representation Learning for Multimodal VAEs», Neurips 2024
[2] Agostini, …, Vogt and **Sutter**, «Weakly-Supervised Multimodal Learning on MIMIC-CXR», ML4H, 2024

MIMIC-CXR: Comparison with Supervised Approaches



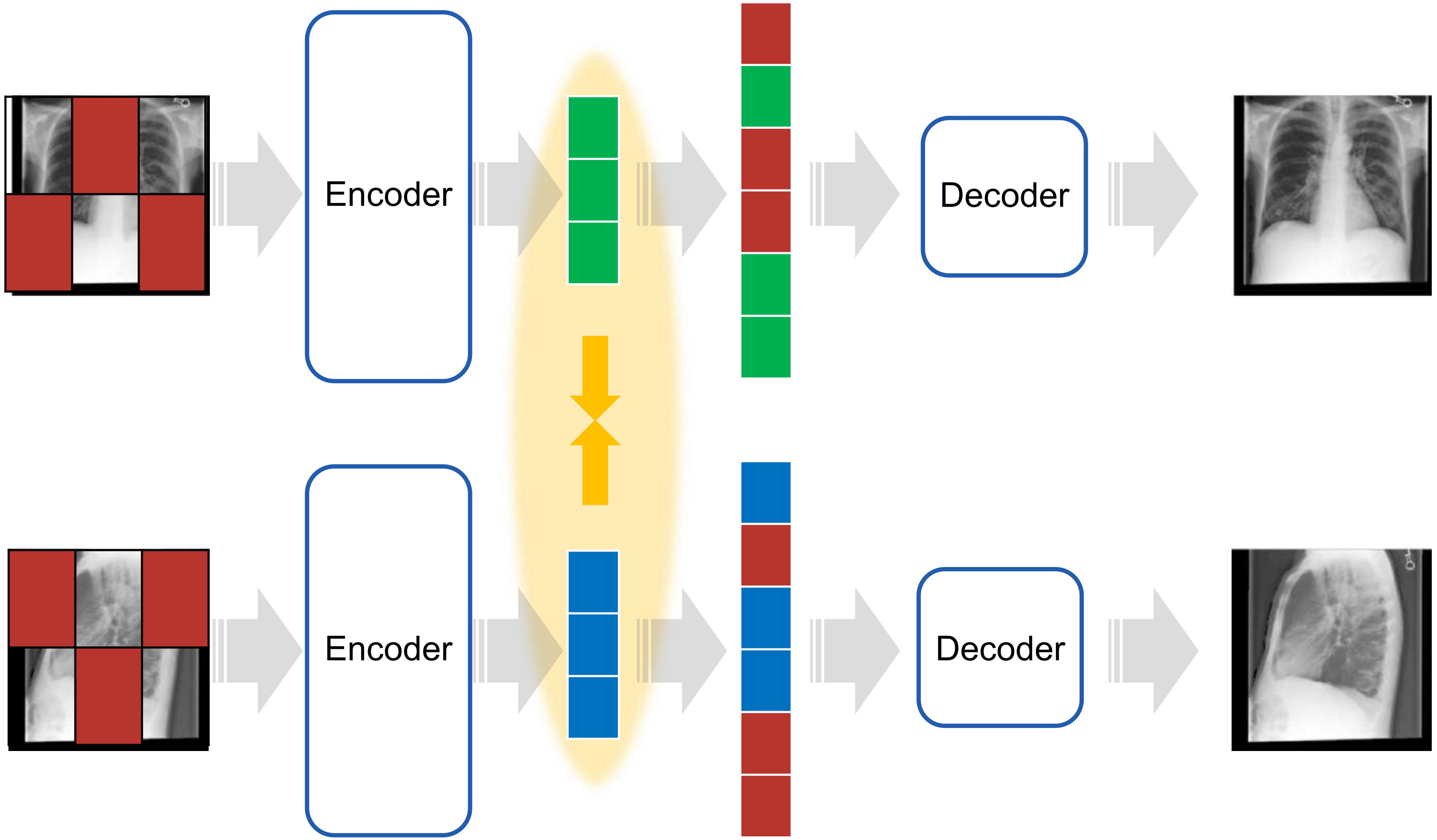
[1] **Sutter** et al., «Unity by Diversity: Improved Representation Learning for Multimodal VAEs», Neurips 2024
[2] Agostini, …, Vogt and **Sutter**, «Weakly-Supervised Multimodal Learning on MIMIC-CXR», ML4H, 2024

Masked Autoencoders: A more modern Approach



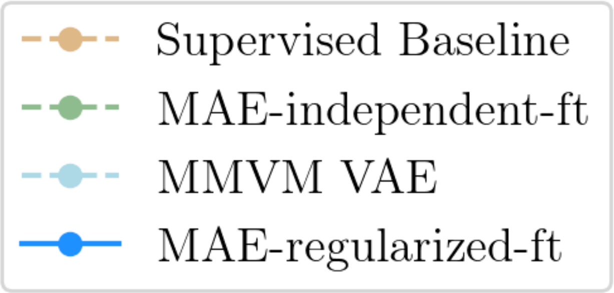
Picture from He et al., «Masked Autoencoders are scalable Vision Learners», CVPR 2022

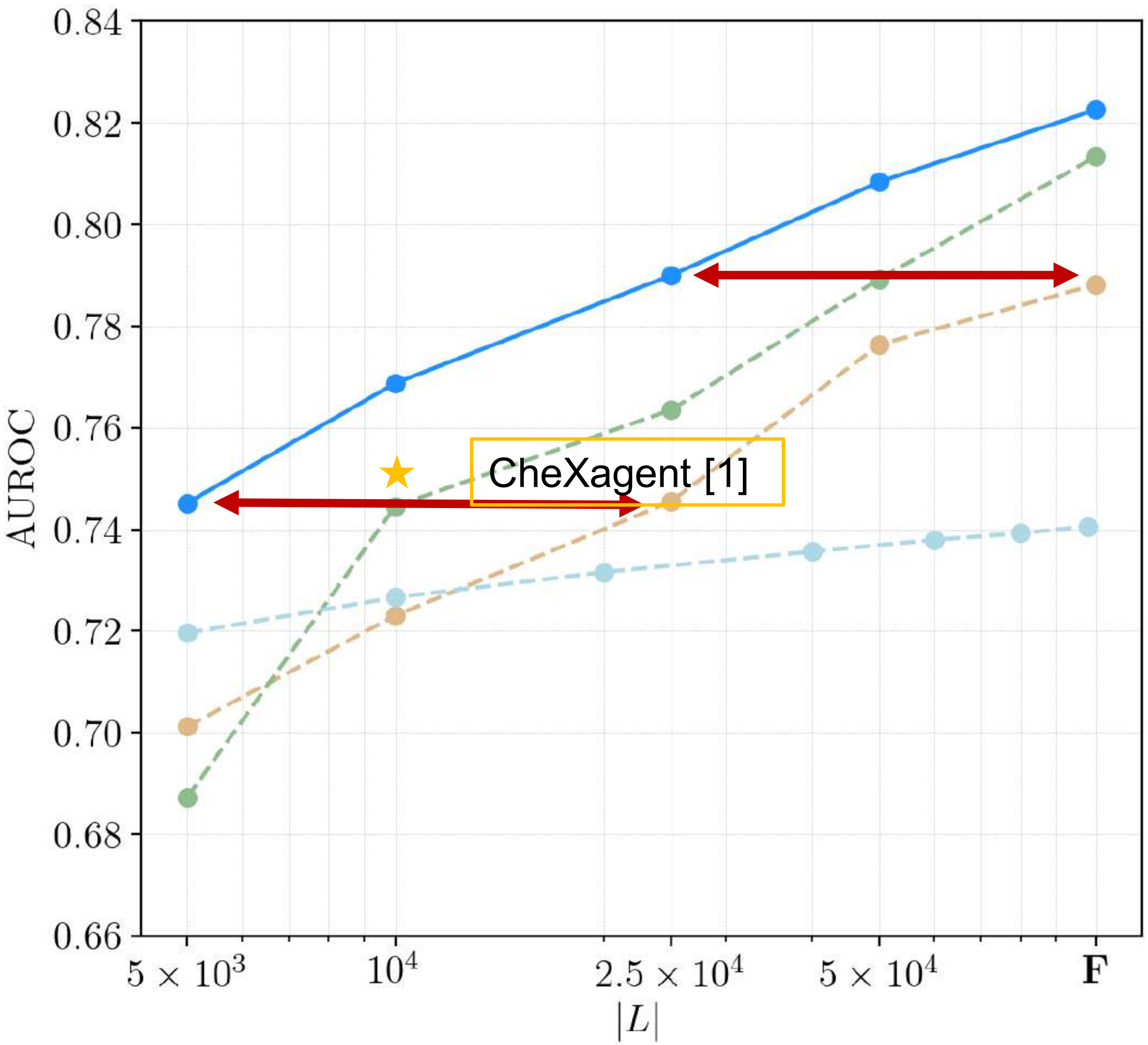
Regularized Masked Autoencoders



[1] Agostini, …, Vogt and **Sutter**, «Leveraging the Structure of Medical Data for Improved Representation Learning», under submission, 2025

Regularized MAE: Results





★ CheXagent [1]

0.62 ★ CLIP [2]

[1] Chen et al., «CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation», arxiv preprint, 2024

[2] Radford et al., «Learning Transferable Visual Models From Natural Language Supervision», ICML, 2021

Conclusion & Future Steps

MIMIC

- Include additional modalities: timeseries, lab values, US , ECG, etc

Multimodal ML

- Novel multimodal objective: strong results on MIMIC-CXR
- Regularization can help improve performance

General

- Multimodal learning is key in applying ML to the medical domain: challenges and opportunities
- Self-supervised learning especially beneficial in the specialized domains
- Ideas from multimodal learning are broadly applicable

ETH zürich

Thomas M. Sutter
Postdoc
thomas.sutter@inf.ethz.ch

ETH Zürich
Medical Data Science
CAB G 37.1
Universitätsstrasse 6
8006 Zürich

https://thomassutter.github.io/

medical data science