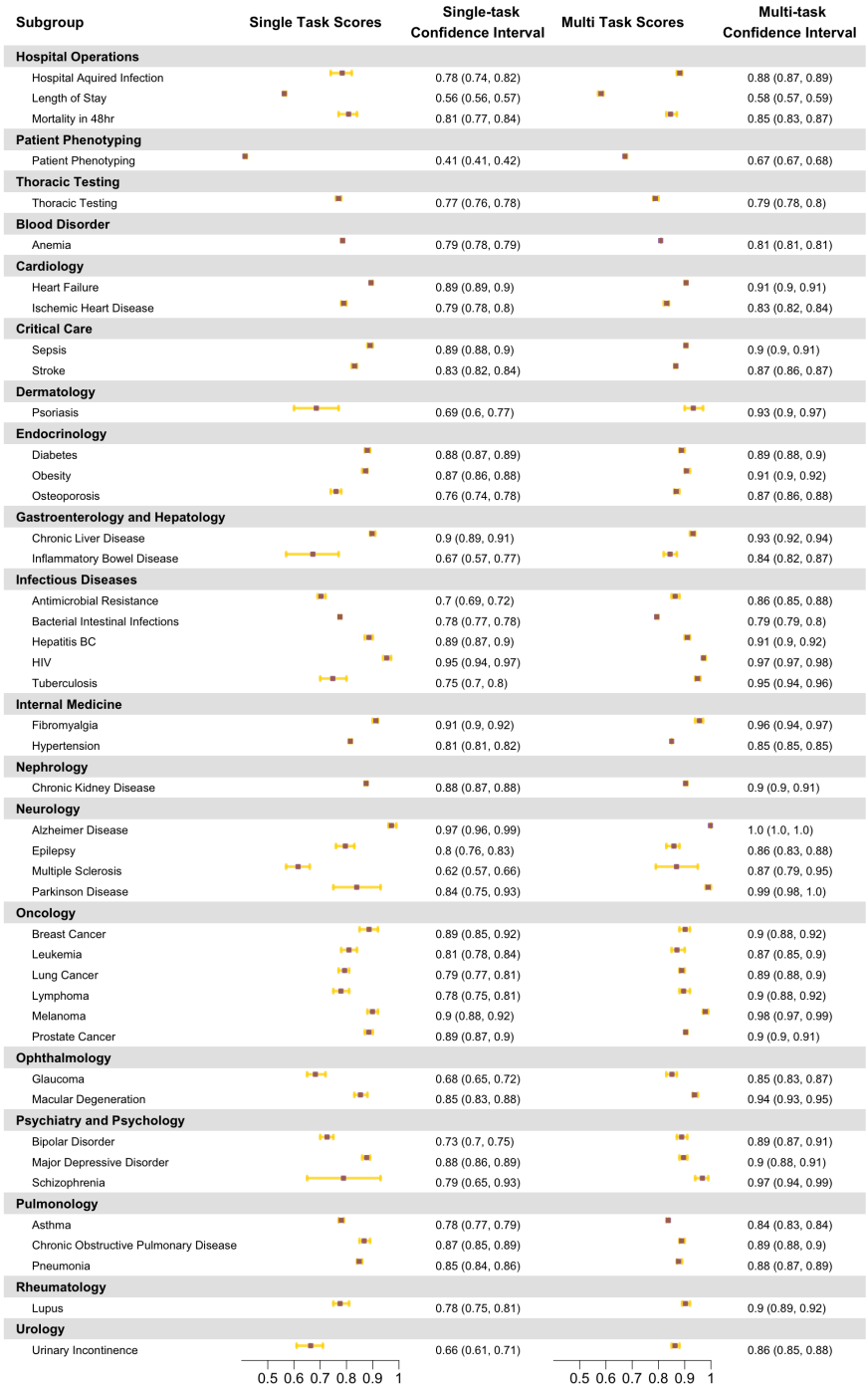
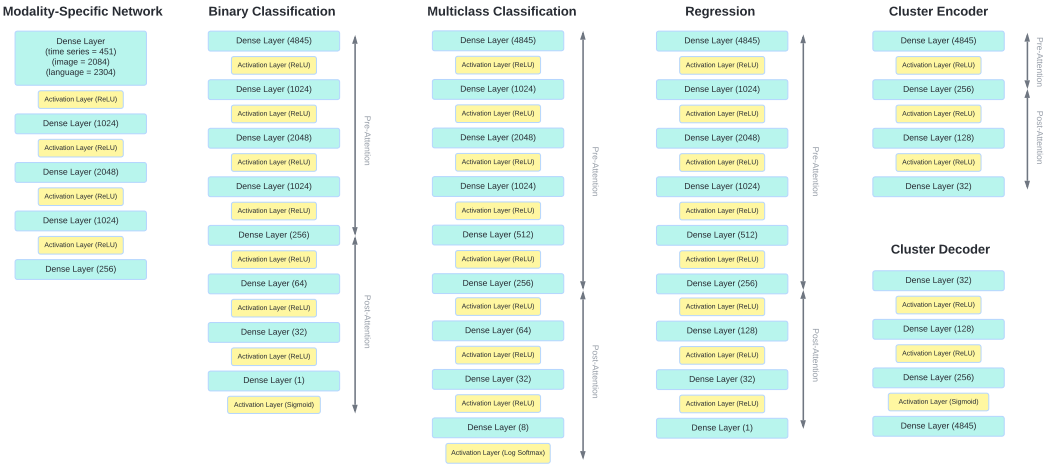
A Appendix / supplemental material

A.1 Computational Results

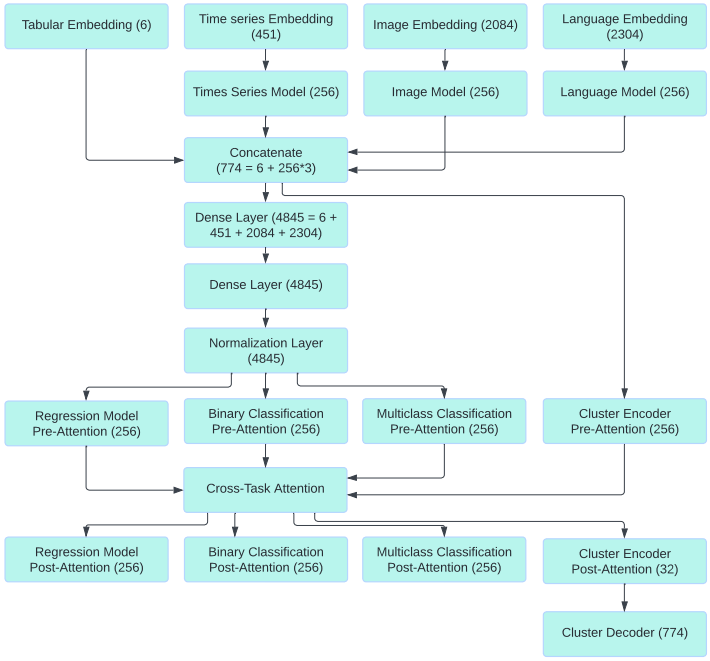


Supplemental Figure A1 Comparison of the Performance of Single-task and Multi-task Models Across Important Healthcare Tasks.

A.2 Architecture Details



Supplemental Figure A2-1. Modality-specific and task-specific network architectures.



Supplemental Figure A2-2. Overall Pipeline of the M3H Architecture.

A.3 Step-by-Step Data Integration and Modeling Procedure

Algorithm A3 End-to-End Data Integration and Modeling Pipeline

Input: Tabular data $X^{tabular}$, Time-series data $X^{time-series}$, Image data X^{vision}, Language data $X^{language}$, Feature extractor f_i of modality i, Outcome vector y_k for task $k \in \mathcal{K}$, $\hat{k} \in \hat{\mathcal{K}}$ indicates a set of task combinations. $\mathcal{L}_{\hat{k}}$ as the aggregated loss function of each task combination $\hat{k}$. $p \in \mathcal{P}$ is the set of hyperparameter combinations. $\epsilon = 10^{-6}$ to avoid numerical precision error during computation,
Output: Trained model and evaluation scores

Step 0 – Data pre-processing and cleaning

- Impute missing values for all modalities, where here x is a generic data entry:

$$x = \begin{cases} 0 & \text{if } x \text{ is numerical or image data} \\ \text{"" (empty string)} & \text{if } x \text{ is text data} \end{cases}$$

- Rescale image size:

$$X^{vision} \leftarrow \text{resize}(X^{vision}, 224 \times 224)$$

Step 1 – Embedding generation of each modality, an example of difference sources with the same modality is EKG notes vs. radiology notes:

$$E_j^i = f_i(X_j^i) \quad \forall i \in \{tabular, \textit{time-series}, vision, language\}, \\ \forall j \in \{\textit{different sources in each modality}\}$$

Step 2 – Concatenate embeddings of all sources of the same modality into a single flattened vector:

$$E^i = \text{vec}(E_1^i, E_2^i, \cdots, E_n^i) \quad \forall i \in \{tabular, \textit{time-series}, vision, language\}$$

Step 3 – Data Normalization

$$E^i = \frac{E^i - \text{mean}(E^i)}{\text{STD}(E^i) + \epsilon} \quad \forall i \in \{tabular, \textit{time-series}, vision, language\}$$

Step 4 – Structure input data with outcomes for a task combination

$$E_{\hat{k}} = \text{vec}(E^{tabular}, E^{\textit{time-series}}, E^{vision}, E^{language}, y_1, y_2, \cdots, y_k)$$

Step 5 – Model Training, Validation and Evaluation

For task combination $\hat{k}$ in the set of all prediction tasks $\hat{\mathcal{K}}$:

- Split data into train and test datasets with fixed seed.

$$E_{\hat{k}}^{train}, E_{\hat{k}}^{test}, y_{\hat{k}}^{train}, y_{\hat{k}}^{test} \leftarrow \textit{train\_test\_split}(E_{\hat{k}})$$

- Perform 5-fold cross-validation with grid search to select the best parameter combination $p^* \in \mathcal{P}$ on the training data that has the best cumulative performance across all tasks inside the task combination $\hat{k}$.

$$M3H_{\hat{k}}^* \leftarrow \underset{M3H_p \forall p \in \mathcal{P}}{\text{argmin}} \mathcal{L}_{\hat{k}}(M3H_p(E_{\hat{k}}^{train}), y_{\hat{k}}^{train})$$

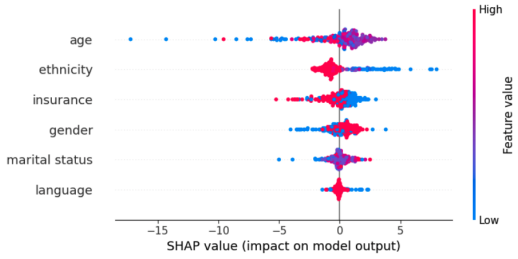
- Evaluate the optimal M3H model on the test set data:

$$\text{test\_set\_score} = M3H_{\hat{k}}^*(E_{\hat{k}}^{test}, y_{\hat{k}}^{test})$$

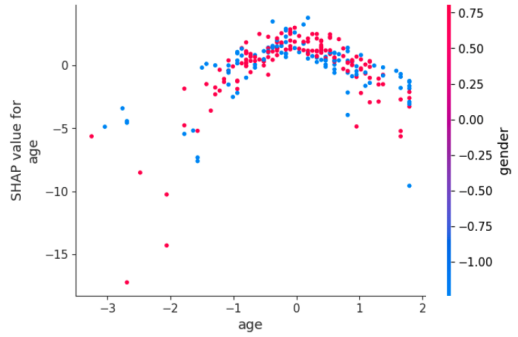
For each potential number of task combinations (i.e., single task = 1, 3-combined multitask = 3) and each task k, report the best model performance for each task.

A.4 Explainability of Input Space: by SHAP

We demonstrate below that by using SHAP values, we can effectively understand the magnitude and directionality of each input clinical variable's contribution to the outcome prediction and thus provide actionable insights for the physicians. Specifically, we analyze an M3H-framework- trained multi-task model between diabetes and heart failure and study the effects of tabular features on diabetes outcomes. We sampled 100 patients and studied their mean-standard deviation normalized tabular features using two types of analysis: feature importance and feature interaction. We observe that patients with lower age are less likely to have diabetes (blue dots for age have mostly negative SHAP values).



Supplemental Figure A4-1. SHAP feature importance plot: each dot indicates a single sample among the 100 test set samples. Higher values of the feature are indicated in red, and lower values in blue. The most important feature is ranked at the top, followed by other features. A higher SHAP value (right-hand side of the axis) indicates a higher likelihood of a positive outcome (has diabetes), and a lower SHAP value indicates a negative outcome (does not have diabetes).



Supplemental Figure A4-2. The SHAP interaction plot demonstrates the nonlinear interactions between features on the outcome prediction captured by the M3H model. Age impacts the risk of diabetes differently depending on the patient's gender.

A.5 Characteristics of HAIM-MIMIC-MM

A.5.1 Limitations:

HAIM-MIMIC-IV was developed from the MIMIC-IV database, and several inherent biases and limitations should be addressed. The cohort is collected from a single-care hospital in Boston and focuses on intensive-care unit patients. This could potentially restrict the demographics and clinical conditions of the patients to this specific geographical location and hospital setting. We also note that MIMIC-IV has recording errors, missing values, and other inconsistencies that are universal to all medical datasets and could pose a challenge for model development.

A.5.2 Embedding Dimensionality and Corresponding Clinical Variables:

The embeddings used as input data for M3H come from the multimodal database HAIM-MIMIC-MM, where the dimensionality of the features is explained and summarized in the paper's original supplemental tables 1 and 2, which are included below for reference. The size of time-series embedding is computed as the number of raw features multiplied by 11 unique features extracted: maximum, minimum, mean, variance, average piece-wise change over time, average absolute piece-wise change over time, maximum absolute piece-wise change over time, sum of absolute piece-wise change over time, change from end-beginning magnitude, number of peaks, and slope of the original time series sequence. There are three categories: chart event ($9 \times 11 = 99$ features), lab event ($22 \times 11 = 242$ features), and procedure event ($10 \times 11 = 110$ features). The size of note embedding comes from the output shape of the pre-trained model ClinicalBERT, which is 768. Similarly, the size of vision embeddings comes from the output shape of the pre-trained model Densenet121-res224-chex, which is 1024 (the dimension of the second to last layer of the model), and 18 (the output/last layer dimension).

A.5.3 Missing Data:

We also include here a table of the missing value distribution of the HAIM-MIMIC-MM dataset reported in the original paper (originally Supplemental Table 3) and how it was handled in that integration procedure.

#	Chart events	Laboratory events	Procedure events
1	Heart rate	Glucose	Foley Catheter
2	Non-invasive systolic blood pressure	Potassium	PICC Line
3	Non-invasive blood diastolic pressure	Sodium	Intubation
4	Non-invasive nominal blood pressure	Chloride	Peritoneal dialysis
5	Respiratory rate	Creatinine	Bronchoscopy
6	O_2 saturation by pulse oximetry	Urea nitrogen	EEG
7	Verbal GCS response	Bicarbonate	Dialysis CRRT
8	Eye opening GCS response	Anion gap	Dialysis catheter
9	Motor GCS response	Hemoglobin	Chest tube removed
10		Hematocrit	Hemodialysis
11		Magnesium	
12		Platelet count	
13		Phosphate	
14		White Blood Cells	
15		Total calcium	
16		MCH	
17		Red Blood Cells	
18		MCHC	
19		MCV	
20		RDW	
21		Platelet count	
22		Neutrophils	
23		Vancomycin	

Supplemental Table A5-1. Patient signals in MIMIC-IV-MM by type of event used as time-series for embedding extraction. Nine time-dependent signals were derived from procedures, twenty-three were derived from laboratories, and eight were derived from information included in the patient chart. CRRT=Continuous renal replacement therapy, EEG=Electroencephalogram, GCS=Glasgow Coma Scale, MCH=Mean corpuscular hemoglobin, MCHC=Mean corpuscular hemoglobin concentration, PICC=Peripherally inserted central catheter, RDW=Red blood cell distribution width.

#	Data Modalities	#	Data Sources
1	Tabular	1	Demographics (E_{de})
2	Time-series	2	Chart events (E_{ce})
		3	Laboratory events (E_{le})
		4	Procedure events (E_{pe})
3	Text	5	Radiological notes (E_{radn})
		6	Electrocardiogram notes (E_{ecgn})
		7	Echocardiogram notes (E_{econ})
4	Images	8	Visual probabilities (E_{vp})
		9	Visual dense-layer feature (E_{vd})
		10	Aggregated visual probabilities (E_{vmp})
		11	Aggregated visual dense-layer features (E_{vmd})

Supplemental Table A5-2. List of different data modalities and data sources used to test the HAIM framework based on the MIMIC-IV-MM database. There are a total of four data modalities and eleven data sources. All data sources correspond to only one data modality. Thus, a model trained on a single data modality can have as little as 1 data source and many as 4 different data sources (of the same kind) as inputs. Double, triple and quadruple modality models can have a number of data sources ranging from [2 to 7], [3 to 9] and [4 to 11], respectively.

Feature Name	Missing %	Source	Handling
anchor_age	0.0	Demographics	N/A
gender_int	0.0	Demographics	N/A
ethnicity_int	0.0	Demographics	N/A
marital_status_int	0.0	Demographics	N/A
language_int	0.0	Demographics	N/A
insurance_int	0.0	Demographics	N/A
Foley Catheter	82.6	Procedure	Fill with 0
PICC Line	63.7	Procedure	Fill with 0
Intubation	75.3	Procedure	Fill with 0
Peritoneal Dialysis	99.7	Procedure	Fill with 0
Bronchoscopy	81.5	Procedure	Fill with 0
EEG	91.5	Procedure	Fill with 0
Dialysis - CRRT	93.1	Procedure	Fill with 0
Dialysis Catheter	88.9	Procedure	Fill with 0
Chest Tube Removed	93.1	Procedure	Fill with 0
Hemodialysis	92.9	Procedure	Fill with 0
Glucose	4.4	Lab	Fill with 0
Sodium	4.7	Lab	Fill with 0
Potassium	4.7	Lab	Fill with 0
Chloride	4.7	Lab	Fill with 0
Creatinine	4.7	Lab	Fill with 0
Urea Nitrogen	4.7	Lab	Fill with 0
Bicarbonate	4.7	Lab	Fill with 0
Anion Gap	4.7	Lab	Fill with 0
Hemoglobin	4.7	Lab	Fill with 0
Hematocrit	4.8	Lab	Fill with 0
Magnesium	5.4	Lab	Fill with 0
Platelet Count	9.8	Lab	Fill with 0

Feature Name	Missing %	Source	Handling
Phosphate	6.0	Lab	Fill with 0
White Blood Cells	4.9	Lab	Fill with 0
Calcium, Total	6.0	Lab	Fill with 0
MCH	4.9	Lab	Fill with 0
Red Blood Cells	4.9	Lab	Fill with 0
MCHC	4.9	Lab	Fill with 0
MCV	4.9	Lab	Fill with 0
RDW	4.9	Lab	Fill with 0
Neutrophils	36.9	Lab	Fill with 0
Vancomycin	60.0	Lab	Fill with 0
Heart Rate	19.5	Chart	Fill with 0
Non-Invasive Blood Pressure systolic	23.4	Chart	Fill with 0
Non-Invasive Blood Pressure diastolic	23.4	Chart	Fill with 0
Non-Invasive Blood Pressure mean	23.3	Chart	Fill with 0
Respiratory Rate	19.5	Chart	Fill with 0
O_2 saturation pulse oximetry	19.6	Chart	Fill with 0
GCS - Verbal Response	20.8	Chart	Fill with 0
GCS - Eye Opening	20.7	Chart	Fill with 0
GCS - Motor Response	20.8	Chart	Fill with 0
Electrocardiogram Notes	11.2	Notes	Empty String
Echocardiogram Notes	30.5	Notes	Empty String
Radiology Notes	0.1	Notes	Empty String

Supplemental Table A5-3. List of missing data percentages by individual variables and handling strategy. Individual variables (i.e., feature name) within key MIMIC-IV-MM data source groups are shown. The strategy for missing value handling used in our tests is as follows: 1) We exclude patients with no available X-rays from our selection cohort; 2) Time-series features are imputed with 0 if there is no measurement at any timestamp; 3) Text embeddings are generated from an empty string if there is no note available; 4) There were no missing values for demographics data.

A.6 Multitask Comparison

We implemented three methods using a universal problem setting of N tasks with feature dimension of d, with input features $X = \{x_j\}_{j=1}^{N}$ where $x_j \in \mathbb{R}^d$. We do not include reshaping operations or the batch size dimension in the description to capture only the mathematical essence of the implementations.

Multi-head attention:

- Initialize linear transformation matrices:
  - $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$ as query, key, value transformations
  - $W^O \in \mathbb{R}^{d \times d}$ as the output transformation
  - $H = 4$ as the number of heads
  - $d_H = d/H$ as the dimension per head
- Apply linear transformation and projection on input features:
  - $Q = XW^Q, K = XW^K, V = XW^V$
- Scaled dot-product to obtain attention weight ($\sqrt{d_h}$ is used to stabilize gradient):
  - $A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)$
- Apply attention weights to obtain output:
  - $O = (AV)W^O$

Cross-stitch:

- Initialize task interaction matrix:
  - $\{T_{ij}\}_{i \neq j}^{1:N}$ where $T_{ij} \in \mathbb{R}^{2 \times 2}$
- Apply interaction matrix:
  - $z_{ij} = T_{ij} \cdot [x_i, x_j] \quad \forall (i, j)$
- Aggregation:
  - $z_i = \sum_{j=1}^{N} z_{ij} \quad \forall\, i = 1, \ldots, N$
- Output learned features:
  - $\{z_i\}_{i=1}^{N} \quad \forall\, i = 1, \ldots, N$

  For n tasks, this requires $\frac{n(n-1)}{2}$ weight matrices of size 2×2.

Multilinear relationship network (MRN):

- Initialize linear transformation matrices:
  - $\{T_{ij}\}_{i,j=1}^{N}$ where $T_{ij} \in \mathbb{R}^{d \times d}$
- Apply linear transformation and projection on input features:
  - $z_{ij} = T_{ij} \cdot [x_i, x_j] \quad \forall (i, j)$
- Aggregation:
  - $z_i = \sum_{j=1}^{N} z_{ij} \quad \forall\, i = 1, \ldots, N$
- Output learned features:
  - $\{z_i\}_{i=1}^{N} \quad \forall\, i = 1, \ldots, N$

  For n tasks, this requires $\frac{n^2}{2}$ weight matrices of size $d \times d$.

Specifically, we conduct experiments in the original dataset on 10 different combinations of multi-tasks that comprehensively evaluate multitask strategies across all four types of machine learning problem classes. The choice of diabetes and heart failure is arbitrary.

- Length of stay (regression), patient phenotyping (clustering)

- Length of stay (regression), thoracic testing (multiclass classification)
- Thoracic testing (multiclass classification), patient phenotyping (clustering)
- Diabetes (binary classification), length of stay (regression)
- Diabetes (binary classification), patient phenotyping (clustering)
- Diabetes (binary classification), thoracic testing (multiclass classification)
- Heart failure (binary classification), length of stay (regression)
- Heart failure (binary classification), patient phenotyping (clustering)
- Heart failure (binary classification), thoracic testing (multiclass classification)
- Diabetes (binary classification), Heart failure (binary classification)

We observe that cross-task attention has a clear advantage in the majority of the cases across all three strategies, with cross-stitch being a close competitor in these 2-tasks experiments (but with qualitative disadvantages discussed below).

Machine Learning Problem Class	**Cross-task (M3H)**	**Multi-Head Attention**	**Multilinear Relationship Network**	**Cross Stitch**
Regression	0.567	0.562 (-0.88%)	0.431 (-23.99%)	0.565 (-0.35%)
Clustering	0.405	0.521 (+28.64%)	0.176 (-56.54%)	0.487 (+20.25%)
Multiclass	0.755	0.715 (-5.30%)	0.595 (-21.19%)	0.755 (+0%)
Binary (diabetes)	0.873	0.824 (-5.61%)	0.873 (+0%)	0.869 (-0.46%)
Binary (heart failure)	0.881	0.864 (-1.93%)	0.896 (+1.7%)	0.888 (+0.79%)

Supplemental Table A6. Comparison of machine learning problem classes across different models. The values represent performance metrics and percentage differences from the baseline (Cross-task M3H).

Beyond the quantitative advantage of the proposed cross-task framework, we would also like to emphasize the qualitative advantage of the chosen framework over existing methods:

- **Interpretability**: Available multimodal multi-task foundation models heavily rely on complex architectures, for example, with repeated use of multi-head attention mechanisms tens or hundreds of times to achieve good performance guarantees. Even with known visualization efforts to interpret these architectures, in practice, these attention weights are almost very often not interpretable and non-sensible. This is why we opted for such a model structure design. As reviewer 2 later correctly pointed out, the existing style of complex architecture makes it very difficult to obtain clinician trust in hospital settings precisely because of such lack of interpretability. Instead, in our case, we apply a single cross-task attention with one single channel and a clean 2D attention weight to explicitly model how self-attention and cross-attention interact. Such design allows for future analysis of interpretability a lot more easily.
- **Lightweight design for deployment**: Existing architectures, such as Google's Med-PaLM 2 (released March 2023), contain 540 billion parameters and can be estimated usually to need months to train with commercial-grade GPUs (such as Nvidia Volta V100) with heavy RAM memory requirements (at least 1000GB if not parallelized). Although lighter-weight versions of these models exist, they remain in the billion-level parameters and pose a significant implementation challenge for hospitals if they wish to host in-house models for data privacy reasons. M3H, on the other hand, can be offered as a much lighter solution to avoid these issues.

Similarly, all three of the compared multiclass methods require significantly more complex network structures. Multi-head model (in our case with 4 heads) requires 4 additional channels to integrate the data from separate heads; cross-stitch models would require significantly more weight matrices as the number of co-learned tasks increases, MRN models will require even more parameters, as they require a linear transformation of each combination of task pairs.

A.7 Loss Function Definition

Binary cross entropy loss (Binary classification):
Given $x \in \mathbb{R}^d$ as an input feature of dimension d, $y \in \{0,1\}^d$ as the binary outcomes, $\hat{y} = \sigma(w^T x + b)$ is the predicted outcome from the M3H framework. Here $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{binary}}(y, \hat{y}) = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y}))$.

Negative log-likelihood loss (Multiclass classification):
Given $x \in \mathbb{R}^d$ as an input feature of dimension d, $y \in \{1, 2, ..., K\}^d$ as the multiclass outcomes from K classes, $\hat{y} = \sigma(w^T x + b)$ is the predicted outcome from the M3H framework. Here: $\sigma(z) = z - \log\left(\sum_{k=1}^{K} e^{z_k}\right)$ is the log-softmax function, w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{multiclass}}(y, \hat{y}) = -\log(\hat{y})$.

Mean absolute error (Regression):
Given $x \in \mathbb{R}^d$ as an input feature of dimension d, $y \in \mathbb{R}^d$ as the regression outcomes, $\hat{y} = w^T x + b$ is the predicted outcome from the M3H framework. Here w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{regression}}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|$.

Mean squared error (Clustering):
Given $x \in \mathbb{R}^d$ as an encoder input of dimension d, $\hat{x} \in \mathbb{R}^d$ as the decoder output of the same dimension, here w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{clustering}}(x, \hat{x}) = \frac{1}{d}\sum_{i=1}^{d} (\hat{x}_i - x_i)^2$.

Contrastive Learning:
This learning aims to project embeddings of different modalities into the same embedding space by contrasting positive pairs (modalities from the same samples) and negative pairs (modalities from dissimilar samples). In the M3H framework, because of the small dimension of the tabular features (6) in comparison to the rest of the three modalities, we only apply contrastive learning among time series, vision, and language data inputs. The formulation is as follows:

Given $\hat{N}$ as the number of permutations between the N samples' three modalities (or N choose 2), and given E_i and E_j as pairs of embedding vectors from different modalities, y_i as the label for the pair of (i, j), where they are either from the same sample (1) or different samples (0). We define a positive margin $p = 0$, and a negative margin $n = 1$. Specifically, for positive pairs, the loss is only computed if $|E_i - E_j| > p$, which aims to decrease positive pairs' distance to 0, and for negative pairs, we only compute the loss when $|E_i - E_j| < n$, which aims to push the distance to be close to 1. The contrastive loss is computed as follows: $L = \frac{1}{N}\sum_{i=1}^{N} \left(y_i \max(0, |E_i - E_j| - p)^2 + (1-y_i)\max(0, n - |E_i - E_j|)^2\right)$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have provided detailed outline of how experiments are conducted and the improvements of our model in comparison to the nominal single-task models both in the introduction and abstract. We have also highlighted key findings and technical novelties introduced in the later sections as well.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitation section at the end of the paper detailing the several possibilities for limitations, ranging from data, to inclusion of other tasks. We have also provided a practical implication section to reflect rigorously how the framework should be adopted in new data and system settings, and what are the potential remedies to deal of potential challenges.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: Our paper is focused on model architecture design and thus does not have theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper clearly outlined the architectures, parameters, data cohort availability, processing steps to ensure that all necessary data is needed for the reproduction of the computational results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Unfortuantely due to privacy reasons we are unable to release the code. But readers are encouraged to contact the authors to its access.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper outlined all the necessary details for where to obtain the data cohort, how to preprocess the dataset, how to split the train, validation, and test sets, the hyperparameters chosen in the paper in order to reproduce all the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars and confidence intervals are provided for the main computational results in the supplemental figure. Details on how these results are computed are also included in the methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] ,

   Justification: The paper's section on practical implementation details the computational resources needed to run the model, as well as an estimated time to run on a local computer. This is also accompanied by the operating system details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] ,

   Justification: The content of this paper complies with NeuRIPS code of ethics, and was conducted with the hopes to advance our understanding of medicine to better patient care.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In the implication to practice section, we discuss thoroughly both the negative and positive implications of the introduced model and its impact if applied to hospital systems.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: In the practical implementation section, we discuss how the model should be validated and evaluated prior to its implementation. This includes safeguards against for example data perturbations to ensure that the model does not negatively impact patient outcome predictions.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data and models used for this paper are properly introduced, elaborated, and cited by the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes] .

Justification: The provided new model has been thoroughly outlined by the paper with detailed instructions on its training parameters and architectures, data used, as well as evaluations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes] .

Justification: This study is conducted using a licensed, but publicly available dataset of human subjects. The details of this cohort has been thoroughly discussed in the patient cohort section.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: IRB approval was not needed due to the use of a public national database.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.