Promoting cross-modal representations to improve multimodal foundation models for physiological signals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many healthcare applications are inherently multimodal and involve multiple types of physiological signals. As sensors for measuring these signals become more ubiquitous, it is increasingly important to improve machine learning methods that consume multimodal healthcare data. Pretraining foundation models is a promising avenue for success. However, methods for developing foundation models in healthcare are still early in exploration and it is unclear which pretraining strategies are most effective given the diverse set of physiological signals collected. This is in part due to challenges of multimodal learning with health data: data across many patients is difficult to obtain and expensive, and there is a lot of inter-subject variability. Furthermore, modalities are often heterogeneously informative across the downstream tasks of interest. Here, we explore these challenges in the PhysioNet 2018 Challenge dataset collected across 1,985 patients. We used a masked autoencoding objective to pretrain a multimodal model on the dataset. We show that the model learns representations that can be linearly probed for a diverse set of downstream tasks. We hypothesize that cross-modal reconstruction objectives are important for the success of multimodal training as they encourages the model to combine information across modalities. We demonstrate that adding modality drop in the input space improves model performance across downstream tasks. We also show that late-fusion models pretrained with contrastive learning objectives are not as effective as across multiple tasks. Finally, we analyze the representations developed in the model. We show how attention weights become more cross-modal and temporally aligned as a result of our chosen pretraining strategy. The learned embeddings also become more distributed in terms of the modalities that each unit in the model encodes. Taken together, our work demonstrates the utility of multimodal foundation models with health data, even across diverse physiological data sources. We further argue how more explicit means of inducing cross-modality may be valuable additions to any multimodal pretraining strategy.

1 Introduction

Healthcare applications often involve integrating information across many modalities. For instance, to diagnose sleep disorders, physicians may evaluate neural, muscular, and respiratory signals [Ibáñez et al., 2018]. Adding to the complexity, the data used in healthcare spans a wide variety of formats (imaging data, time series, etc) and are collected from sensors placed on many different body locations [Acosta et al., 2022]. Many of these sensors for health data are becoming increasingly prevalent in everyday wearable devices [Jeong et al., 2018, Wu and Luo, 2019, Iqbal et al., 2021]. This technological advance is a promising opportunity for personalized healthcare and improving patient care. Thus, it is more and more important to leverage artificial intelligence to aid the interpretation of health data with heterogenous sensors.

In many settings, artificial intelligence has achieved unprecedented success in the development of multimodal foundation models [Jin et al., 2024, Bordes et al., 2024, Wadekar et al., 2024]. For instance, models can now integrate information across language, vision, audio, and video to solve complex tasks and perform human-like feats of reasoning [Radford et al., 2021, Alayrac et al., 2022, Wu et al., 2023, Lu et al., 2024, Mizrahi et al., 2024]. Multimodal foundation models are pretrained in a self-supervised manner on vast amounts of data to link information across modalities. The representations developed by these models are useful for tasks that require multimodal understanding. After pretraining, these models may be further trained on a downstream task or the representations they produce can be used as is. Pretraining strategies often outperform models trained from scratch on the same tasks and require less labeled data [Jin et al., 2024]. The success of multimodal foundation models in other domains suggests that similar advances can be achieved in healthcare settings.

There are further reasons to believe that health data in particular can benefit from foundation model strategies. Annotated data is limited in health data because clinical expertise is often necessary to create labels. Thus, the label efficiency of pretrained models is very useful in this setting. Furthermore, when considering wearable health devices, it becomes more important to develop models that are size-efficient. If a model pretrained on health data can successfully transfer its representations across many downstream tasks, this can greatly save on memory and runtime costs for wearable devices.

However, working with health data also introduces new types of challenges. Pretraining often consumes large amounts of unlabeled data, but patient privacy concerns limit the amount of large datasets available in this domain [Acosta et al., 2022, Shaik et al., 2023]. In addition, the cost associated with deploying many health sensors can make large-scale data collection prohibitively expensive [Acosta et al., 2022]. Thus, it becomes less clear whether pretraining can be as effective as it is in settings like natural language, where large corpora are more widely available. In health applications, it is also common for certain modalities to vary greatly in their informativeness for different downstream tasks [Krones et al., 2024]. This problem is exacerbated in wearable devices since different sensors may suffer from unequal amounts of noise, perhaps due to weaker contact or interference from other devices [Ates et al., 2022, Canali et al., 2022]. This poses a challenge for developing general purpose models that can be used for diverse tasks.

Here, we investigate these challenges by pretraining a multimodal model in a publicly available dataset with 1,985 patients. We are specifically concerned with time series data collected from physiological signals measured overnight from patients. Our contributions are the following:

- We explore the development of a multimodal foundation model in a dataset of diverse physiological signals: electroencephalography (EEG), electromyography (EMG), electrooculography (EOG), and electrocardiology (ECG). We demonstrate the strength of the learned representations in linear probe experiments on a disparate set of downstream tasks.
- We show how explicitly enforcing cross-modal reconstruction in the pretraining objective improves the quality of the learned representations over standard multimodal MAE. We also show how late-fusion models pretrained with contrastive learning does not effectively transfer across multiple tasks.
- We analyze the learned representations to show that attention weights in the model become increasingly cross-modal under the pretraining objective we use. We also show that individual units in the model become more diversely tuned to the different modalities.

2 Related Work

Pretraining models with self-supervised objectives is a popular and effective strategy in machine learning [Ericsson et al., 2022, Gui et al., 2023]. After pretraining, the parameters of the model can be finetuned for some downstream task. Alternatively, another common approach is to freeze the pretrained model and train a lightweight readout head that uses the learned representations from the model to solve downstream tasks. This approach is especially attractive if efficiency in parameter tuning is a priority. In either case, a pretraining paradigm often outperforms training a model from scratch. Pretraining is especially useful if labeled data in the downstream task is limited as it provide a means for experimenters to define inductive biases on the model representations. Self-supervised strategies span several categories, including generative methods, contrastive learning, and autoencoding [Del Pup and Atzori, 2023, Gui et al., 2023]. We limit our discussion to the latter two in the context of multimodal pretraining.

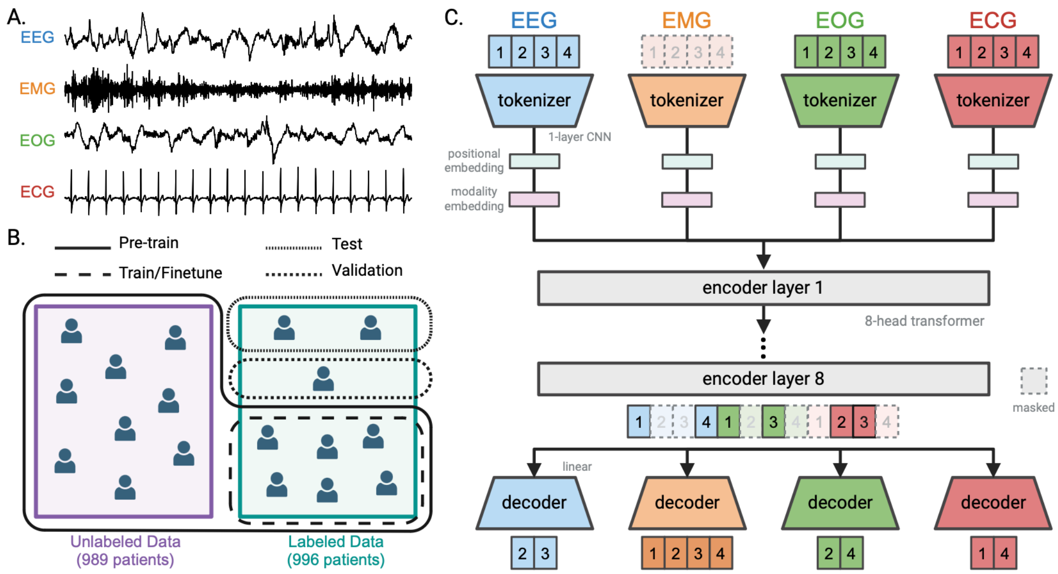

Figure 1: **A.** A 30-second sample from the training dataset. **B.** Data is split by patient identity for each part of the training procedure. The PhysioNet 2018 dataset consists of unlabeled data from 989 patients and labeled data from 996 patients, where each patient contributes 7.7 hours of data on average. The data for pretraining consists of all patients in the unlabeled dataset and 657 patients from the labeled dataset. The data for training and finetuning is drawn from the patients of the labeled dataset that were also used for pretraining. The data for the validation and test are drawn from the remaining patients of the labeled dataset not used for either pretraining or training. **C.** Diagram of the main pretraining strategy we use: multimodal masked autoencoding with modality drop in the input space. Tokenizers are modality-specific.

Contrastive learning is a self-supervised learning framework where models are optimized such that representations of data in positive pairs become more similar while representations of data in negative pairs become more dissimilar [Chen et al., 2020, Purushwalkam and Gupta, 2020]. The definition of positive and negative pairs is crucial to the success of these methods. One way to define these pairs is to construct multiple "views" of a data sample through augmentations, like in the SimCLR algorithm [Chen et al., 2020, Yuan et al., 2021]. Thus, a positive pair of data may be two different augmentations of a single data sample (negative pairs would then be constructed across different data samples). When working with multimodal data, another option is to consider each modality as a distinct view of a data sample. In this case, positive pairs can be constructed by comparing representations across modalities, as in CLIP-style pretraining [Radford et al., 2021, Yuan et al., 2021, Zhang et al., 2022].

Masked autoencoding (MAE) is another popular pretraining strategy. In MAE, random patches of the input are masked, and the model must use the remaining portions of the input to reconstruct the masked portion [He et al., 2022]. This method has been extended to settings with multimodal data, often to combine text data and vision data [Arici et al., 2021, Geng et al., 2022, Bachmann et al., 2022, Zhao et al., 2023, Mizrahi et al., 2024]. To do so, these models combine data across modalities early on so that representations are multimodally fused through layers of the model. The joint embeddings are then used for the MAE task to reconstruct inputs across all modalities. This structure inherently allows for the possibility of cross-modal reconstruction, as information from one modality can be used to reconstruct another. MAE methods can be more compute- and size-efficient due to the fused encoding structure used and the large amounts of data typically dropped out as a result of the masking strategy Bachmann et al. [2022], Mizrahi et al. [2024]. Of the above, the method used for our model is most similar to MultiMAE introduced in Bachmann et al. [2022].

Both of these pretraining strategies have been applied to physiological signals, although examples are sparser than in other domains. We first discuss examples using contrastive learning strategies. Abbaspourazad et al. [2023] uses a large-scale Apple Watch dataset to classify demographics and health information from two modalities: photoplethysmography (PPG) and ECG. Thapa et al. [2024]

used sleep data collected across EEG, EMG, ECG, and EOG sensors for downstream sleep-related classification tasks. Raghu et al. [2022] uses cardiac and blood-related signals to predict mortality rate and pulmonary arterial pressure. Both Abbaspourazad et al. [2023] and Raghu et al. [2022] use a SimCLR-like strategy through data augmentations, while Thapa et al. [2024] uses a CLIP-like strategy and construct data pairs across modalities.

In comparison to contrastive methods, MAE pretraining is less common for multimodal physiological signals. Mathew et al. [2024] uses MAE-pretraining in a model for phonocardiogram (PCG) and ECG data. The data was collected from digital stethoscopes, and the model was finetuned to classify signatures of cardiovascular disease. The closest example to our work is from Liu et al. [2023], where a multimodal transformer model is pretrained on EEG, EMG, and EOG signals with a MultiMAE-like objective. However, this work was more limited in dataset size (100 patients in each pretraining dataset) and focused on one specific downstream task per pretrained model. In our work, we use a larger dataset with 1,985 patients and evaluate how well MultiMAE-pretrained models can perform on diverse downstream tasks. We later will make comparisons with contrastive methods as well.

A focus of our work is in encouraging cross-modal representation learning. This is inspired by works arguing that multimodal learning can be improved by optimizing for cross-modal reconstruction [Kleinman et al., 2023, Hussen Abdelaziz et al., 2020, Hazarika et al., 2022]. While this objective is already present in the original MultiMAE algorithm, a simple way to further encourage cross-modal learning is to randomly drop modalities from the input [Hazarika et al., 2022, Hussen Abdelaziz et al., 2020, Arici et al., 2021, Deldari et al., 2023]. This pressures the model to learn relationships across modalities in order to satisfy the reconstruction task. In the health data field, modality dropout strategies have been used to improve performance in tasks with missing modalities or heterogeneous noise, but they are still limited in their use in a general pretraining strategy. Furthermore, analyses of how representations are shaped by multimodal fusion are largely unexplored. We investigate both these questions in this work.

3 Methods

3.1 Dataset

We use the publicly available PhysioNet 2018 Challenge dataset [Ghassemi et al., 2018]. This dataset consists of physiological signals collected during overnight sleep from 1,985 subjects. On average, each subject contributes 7.7 hours of recording [Ghassemi et al., 2018]. The dataset contains many sensors, but here we focus on EEG, EMG, EOG, and ECG recordings (Figure 1A). For EEG, we use only the F3-M2 differential pair for our main results. We note that the signals from these sensors show distinct characteristics and are not obviously related (Figure 1A).

Patient demographics such as age and gender were also recorded in the dataset. The physiological data comprises of unlabeled data from 989 patients and labeled data from 996 patients. In the labeled set, 30-second contiguous windows were manually annotated by several certified sleep technologists into one of five sleep stages: wakefulness, stage 1, stage 2, stage 3, or rapid eye movement (REM). The same windows were also manually annotated for the presence of arousals (e.g. snores, vocalizations, respiratory effort, leg movement, etc.). Prior literature using this dataset mostly focus on the sleep staging task [Perslev et al., 2019, Banville et al., 2021, Phan et al., 2021], and comparisons to these works are discussed in the Appendix.

To prevent data leakage, data is split over patient identity. The pretraining dataset is comprised of of all 989 patients in the unlabeled set and 657 patients in the labeled set. The training/finetuning dataset for downstream tasks is comprised of the 657 patients in the labeled set that were used for pretraining (that is, the training/finetuning dataset is a subset of the pretraining dataset). The validation set and test set are constructed from the remaining 117 and 219 patients of the labeled set, respectively, and are not seen in either pretraining or training/finetuning. The validation set is used to select hyperparameters of the model and the test set is used for the evaluation scores reported in the results. A visualization of these data splits are in Figure 1B.

The signals are preprocessed with an anti-aliasing bandpass FIR filter then downsampled from 200 Hz to 100 Hz using decimation by 2. Specifically, EEG and EOG signals were filtered to 0.1-30 Hz [Feng et al., 2021, Satapathy et al., 2024]. EMG and ECG signals were filtered to 0.1-70 Hz [Burns

Table 1: *Balanced accuracy with linear probe evaluation: unimodal vs multimodal.* All models are pretrained before the encoder is frozen and representations are linearly probed for each task. We show the test balancy accuracy for a random guess ("Random"), for models trained entirely from scratch ("Scratch"), and for models pretrained and then linearly probed for the task ("Pretrained"). Note that "Pretrained-All' is a multimodal model pretrained with MultiMAE and input modality drop. Mean score and standard deviation for the three tasks are shown in columns. We additionally define an aggregate score which gives the average score over all tasks, normalized by the corresponding chance performance value (a score of 0 would indicate no improvement from chance). 500 patients are used in the training set for task finetuning. 5 random seeds are used in each training/finetuning stage. Asterisks indicate the best-performing unimodal model for each task.

		Sleep	Age	Arousal	Aggregate
Random	–	0.2	0.5	0.5	0.0
Scratch	EEG	0.717 ± 0.003	0.641 ± 0.004	0.568 ± 0.093	1.0 ± 0.101
	EMG	0.461 ± 0.004	0.55 ± 0.006	0.538 ± 0.074	0.494 ± 0.076
	EOG	0.697 ± 0.006	0.626 ± 0.006	0.56 ± 0.082	0.952 ± 0.084
	ECG	0.279 ± 0.006	0.605 ± 0.022	0.516 ± 0.038	0.213 ± 0.025
	All	0.737 ± 0.003	0.626 ± 0.018	0.595 ± 0.013	1.042 ± 0.009
Pretrained	EEG	**0.745 ± 0.001***	0.662 ± 0.001	0.604 ± 0.093	1.085 ± 0.106
	EMG	0.442 ± 0.001	0.615 ± 0.003	0.533 ± 0.048	0.502 ± 0.052
	EOG	0.727 ± 0.001	0.653 ± 0.003	0.636 ± 0.071*	1.07 ± 0.078
	ECG	0.339 ± 0.002	0.703 ± 0.002*	0.526 ± 0.04	0.385 ± 0.042
	All	0.744 ± 0.001	**0.719 ± 0.002**	**0.637 ± 0.081**	**1.144 ± 0.09**

et al., 2007, Feng et al., 2021, Satapathy et al., 2024]. All signals are then resampled to 100 Hz. We use 30-second samples of data for pretraining and for the downstream classification tasks.

Three tasks are constructed from this dataset: (1) sleep scoring, (2) age classification, and (3) arousal identification. Sleep scoring is a 5-way classification problem. Both arousal and age will be treated as a binary classification problem. In the age classification task, we aim to identify whether a patient's age is under 55 (the mean age) or not.

3.2 Model architecture

Our model architecture is based off that of the vision transformer [Alexey, 2020]. Modality-specific tokenizer layers are followed by fused encoding layers, so that multimodal information is fused early on (Figure 1C). The input to the model is a 30 second time series from multiple sensors sampled at 100 Hz. We divide each time series into 30 chunks that are one second each. These chunks are then fed to the tokenizer layers. Tokenizers are trained for each modality and consist of one convolutional layer and one linear layer. Specifically, each signal chunk first passes through a 1D convolutional layer (with 64 channels and kernel size of 21) before a max pooling operation. Then, a linear layer projects each token into a 512-dimensional embedding space. This is followed by layer normalization to ensure signals from all modalities have comparable scales. Given 1,985 patients with an average of 7.7 hours of recording time each, the total dataset size is 1,834,140.

To summarize, the output of a tokenizer for one modality is 30 tokens with embedding dimension $D = 512$. Sinusoidal positional embeddings and a learnable modality embedding are then added to each token. Finally, tokens across modalities are fused through concatenation.

This fused vector is then passed to the joint encoding layers, which is comprised of eight transformer layers with multi-head self-attention [Vaswani, 2017] and normalization before attention layers [Xiong et al., 2020]. Each transformer layer has 8 heads, and each layer has a 10% dropout rate during training over attention weights and projection weights.

3.3 Pretraining objectives

We use a multimodal masked autoencoding (MAE) objective similar to MultiMAE from Bachmann et al. [2022]. As mentioned above, tokens across all modality tokenizers are fused via concatenation. In MultiMAE, a fixed portion of these tokens are masked at uniform and dropped from the fused vector. We use a 70% masking rate (see Appendix for how the mask rate was selected). The remaining unmasked tokens are passed into the encoder and processed. To prepare the input for the decoder layers, the tokens that are output from the encoder are then interleaved with learnable mask tokens. Values in the mask token are initialized from $\mathcal{N}(0, 0.02)$ with truncation at $[-2, 2]$. These learnable mask tokens act as placeholders for the signal to be reconstructed (i.e., the dropped tokens). Mask tokens are inserted in the location of the previously dropped tokens. Positional information is preserved by adding the appropriate positional embedding to the newly interleaved mask tokens.

A decoder is trained for each modality to reconstruct the original signal. Each decoder consists of a cross-attention layer and a transformer layer before a linear projection. The input into each modality decoder is the subset of tokens from the encoder output that corresponds to that modality. Cross-modal reconstruction is enabled through the cross-attention layer, where the input is the query and the entire encoder output is passed as keys/values. The linear layer projects each token from the embedding dimension (512) to the original signal dimension (100). The loss is calculated only over the reconstructed signal corresponding to the dropped tokens.

To encourage additional cross-modal interactions, we also use input modality drop during pretraining (Figure 1B) [Hazarika et al., 2022, Hussen Abdelaziz et al., 2020, Arici et al., 2021, Deldari et al., 2023]. In every batch, one randomly chosen modality is completely dropped on top of the typical MultiMAE uniform masking over tokens.

In later experiments we will make comparisons with contrastive learning objectives, resulting in modifications to the pretraining loss and the model architecture. In this case, the model will be converted to a late fusion structure that is typical for models trained with contrastive objectives. Further details can be found in the corresponding results section (§4.3) and Appendix F.

3.4 Finetuning

We are most interested in understanding how well representations learned by the pretrained model can transfer to multiple tasks. As such, after pretraining, we discard the decoder and freeze the encoder. The output of the encoder is layer normalized and average pooled over the token dimension. This 512-dimensional vector is then passed to a linear classification head. A classification head is trained for each of the downstream tasks with weighted cross entropy loss to account for class imbalance. This is most relevant for the arousal detection task, where arousal events are extremely rare (2.7% of data samples). Although not the focus of this paper, we also conduct full finetuning experiments where both encoder and classifier parameters are trained (Appendix E).

3.5 Optimization

Models are pretrained for 2000 epochs or until a fixed compute time of 10 days is exceeded. For finetuning, models are trained for 200 epochs. Learning rates were scheduled with 10 epochs of linear warmup to 1×10^{-4} and cosine annealing thereafter [Loshchilov and Hutter, 2016]. We used the AdamW optimizer [Loshchilov and Hutter, 2017]. The model checkpoint chosen for evaluation was from the epoch where the lowest validation error was achieved, except in the case of pretraining the MultiMAE model with input modality drop. In this case, the validation error was quite noisy and the most recent checkpoint was chosen instead. Additional details can be found in the Appendix.

4 Experiments

4.1 A pretrained multimodal model develops representations that support a diverse set of tasks in the PhysioNet18 dataset.

We first assess the extent to which pretraining and multimodal learning benefits downstream task performance in this dataset. We evaluate performance on the three tasks when both unimodal and multimodal models are trained from scratch. The balanced accuracy achieved by these models on the test set is shown in Table 1 ("Scratch" rows). In addition to the three tasks, we also define an aggregate score to highlight models that perform well across all tasks. The aggregate score is defined as $\frac{1}{N}\sum_i^N \frac{s_i - r_i}{r_i}$, where s_i is the average test score on task i, r_i is the chance level performance for task i, and $N = 3$ is the total number of tasks. Scores are measured using balanced accuracy. We see that the multimodal model performs overall better than any of the unimodal models (compare

Table 2: *Linear probe evaluation on all three tasks, comparing multimodal pretraining strategies*. All models are pretrained before the encoder is frozen and representations are linearly probed for each task. Mean test score and standard deviation for the three tasks are shown in columns. Aggregate score is defined as in Table 1. 500 patients are used in the training set for task finetuning. 5 random seeds are used in each training/finetuning stage.

Pretraining Strategy	Sleep		Age	
	Balanced Acc.	Cohen Kappa	Balanced Acc.	AUROC
Contrastive CLIP-style (LOO)	0.708 ± 0.0004	0.572 ± 0.001	0.643 ± 0.004	0.705 ± 0.006
Contrastive CLIP-style (Pairwise)	0.703 ± 0.001	0.559 ± 0.001	0.646 ± 0.0003	0.698 ± 0.0003
Contrastive SimCLR-style	0.656 ± 0.001	0.52 ± 0.001	0.624 ± 0.009	0.673 ± 0.015
MultiMAE Only	0.734 ± 0.001	0.618 ± 0.001	0.684 ± 0.001	0.758 ± 0.001
MultiMAE + Input Mod. Drop	**0.744 ± 0.001**	**0.63 ± 0.002**	**0.719 ± 0.002**	**0.785 ± 0.002**

Pretraining Strategy	Arousal		Aggregate Score
	Balanced Acc.	AUROC	
Contrastive CLIP-style (LOO)	**0.71 ± 0.002**	**0.776 ± 0.001**	1.082 ± 0.005
Contrastive CLIP-style (Pairwise)	0.708 ± 0.002	0.772 ± 0.001	1.075 ± 0.002
Contrastive SimCLR-style	0.585 ± 0.048	0.616 ± 0.070	0.900 ± 0.040
MultiMAE Only	0.604 ± 0.089	0.638 ± 0.136	1.082 ± 0.062
MultiMAE + Input Mod. Drop	0.637 ± 0.081	0.677 ± 0.128	**1.144 ± 0.058**

aggregate scores), although its performance on the sleep classification task slightly lags behind the unimodal EEG model.

We next examine the benefits of pretraining the model and transferring the learned representations to each of the downstream tasks. We first pretrain the unimodal models with masked autoencoding. The test scores for these models are shown in Table 1 as well ("Pretrained" rows). Pretraining seems to benefit all models, whether unimodal or multimodal. Interestingly, the pretrained unimodal models reveal that a different modality is most informative for each task: EEG is more effective for sleep staging, ECG for age classification, and EOG for arousal classification.

We then pretrain a multimodal model with MultiMAE and input modality drop. We evaluate this model on the downstream tasks ("Pretrained, All" in Table 1). The multimodal model outperforms the unimodal model in age classification and arousal classification, and performs very similarly to EEG in the sleep staging task (Table 1). We find that the multimodal model performs well in all tasks and achieves a higher aggregate score, despite the imbalance in modality dominance across tasks.

Notably, the improvement in aggregate score obtained by the multimodal model is greater when training data is more limited (Appendix D). Given full-finetuning, though, the differences across models are more minimal (Appendix E).

4.2 Adding input modality drop to MAE pretraining is important for downstream task performance.

We chose our particular pretraining strategy with the hypothesis that encouraging multimodal fusion improves performance in the downstream tasks. We investigate whether this is the case by first testing the importance of using input modality drop (which theoretically should result in more cross-modal learning). We compare task performance to that of a standard MultiMAE strategy, which does not include input modality drop (Table 2). We see that removing input modality drop causes a performance drop in downstream tasks (compare "MAE Only" to "MAE + Input Mod. Drop" in Table 2). In fact, without input modality drop, MultiMAE underperforms the most informative unimodal models across all tasks. Overall, dropping modalities in the input appears to be a simple and effective means to increase performance over standard MultiMAE.

4.3 Late fusion models with contrastive learning objectives are more variable in performance.

Multimodal fusion is additionally encouraged in the MultiMAE model through the early fusion architecture, where representations across modalities are mixed early in the network. We next make comparisons to models with a late fusion structure where representations across modalities are not

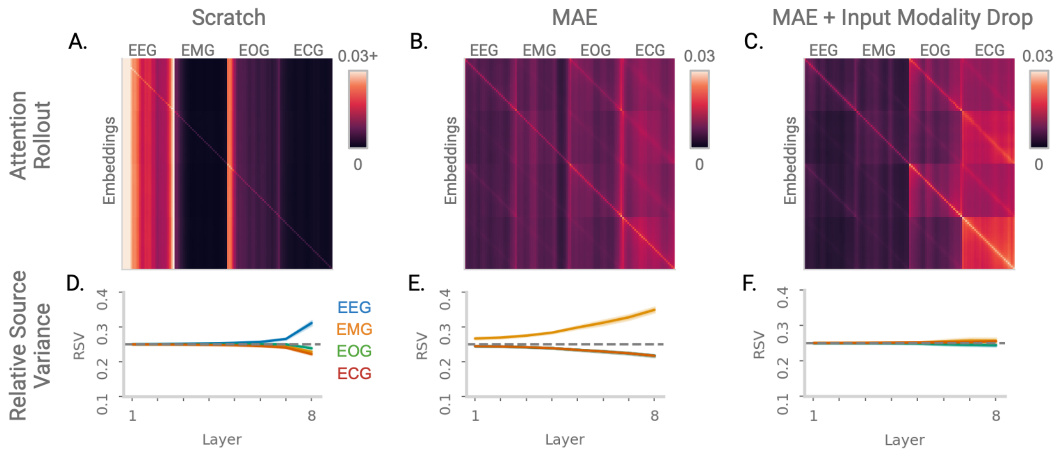

Figure 2: Measures of modality fusion across model representations. **A.** Attention rollout from tokens in the embeddings to tokens in the input. Here, the model is trained from scratch on sleep staging. Values are capped at 0.03 for comparisons with (BC). **B.** As in (A), but for the model pretrained with MAE. **C.** As in (A), but for the model pretrained with MAE and input modality drop. **D.** Relative source variance (RSV) of units across layers of the model in (A) to each of the four modalities. 95% confidence intervals shown, over 512 units in each embedding vector. **EF.** As in (D), but for the models in (B) and (C), respectively.

mixed except in the decoders for downstream tasks (Appendix F). To do so, we pretrain late fusion models with contrastive learning, a common pretraining objective for these types of model.

We first test SimCLR-style multiview contrastive learning [Chen et al., 2020, Purushwalkam and Gupta, 2020], with particular inspiration from Raghu et al. [2022]. We randomly generate augmentations for all input data samples (using the same signal augmentations from Raghu et al. [2022]). Positive pairs are defined as representations from adjacent time windows. We find that the SimCLR-style model underperform standard MultiMAE in all tasks (Table 2). This may indicate that defining desired relationships between of modality embeddings is important for the performance of a contrastive learning model.

We next test CLIP-style pretraining to assess the benefits of using modality contrast in the contrastive learning loss [Radford et al., 2021, Yuan et al., 2021, Zhang et al., 2022]. We will use two objectives defined in Thapa et al. [2024], a previous work in physiological signals that inspired our approach here. Thapa et al. [2024] defined a pairwise loss and a leave-one-out (LOO) loss:

$$l_{ijk}^{pair} = -\log \frac{\exp(\text{sim}(x_k^i, x_k^j)) * \tau}{\sum_{m=1}^{N} \exp(\text{sim}(x_k^i, x_m^j)) * \tau} \qquad l_{ik}^{LOO} = -\log \frac{\exp(\text{sim}(x_k^i, \bar{x}_k^{\neq i})) * \tau}{\sum_{m=1}^{N} \exp(\text{sim}(x_k^i, \bar{x}_m^{\neq i})) * \tau}$$

for modalities i and j, sample k, temperature τ, and modality embedding x. N is the total number of samples, and $\bar{x}_k^{\neq i}$ is the average of representations that are not modality i given data sample k. We find that both CLIP-style models underperform even standard MultiMAE in sleep and age classification (Table 2). Surprisingly, the contrastive model does extremely well in arousal classification. However, in terms of aggregate performance, using MultiMAE with input modality drop is still most effective out of the strategies we tested.

Despite these results, we speculate that developing new formulations of contrastive learning may improve task performance. These methods are highly sensitive to the choice of positive and negative pairs. It may be that contrastive methods in multimodal biosignals require domain-specific design to reach their full potential.

4.4 MAE + input modality drop encourages cross-modal fusion in attention weights and model representations.

Finally, we wanted to understand whether our intuition about cross-modal fusion was indeed reflected in the representations developed by the model. We first examine the attention weights of the model to understand how much each output token from the encoder is influenced by input tokens from each modality. We use a method called attention rollout [Abnar and Zuidema, 2020]. Attention rollout accounts for the effects of the residual layers by defining the attention at layer l as a sum of the raw attention weights and the identity matrix: $A_l = 0.5W_l + 0.5I$ where $W_l = \text{softmax}(Q_l K_l^T)$. Thus, to obtain the attention of the output embedding to the inputs, attention weights are rolled out across model layers: $A_L * A_{L-1} * \cdots * A_2 * A_1$, for L layers in the encoder.

We plot the results of attention rollout first for a multimodal model trained from scratch on sleep-scoring (Figure 2A). The attention matrix develops strong vertical bands, indicating that model embeddings attend to specific tokens without any context-specificity. In this case, EEG and EOG tokens are most dominant. We next plot the attention matrix for a model pretrained with MultiMAE and MultiMAE with input modality drop (Figure 2BC). The attention weights are more evenly spread across the matrix, indicating greater cross-modal attention, although some sparse vertical bands can still be observed. This can be interpreted as greater context-specificity in the attention weights. We also observe an additional effect from both MultiMAE models where attention weights become more temporally aligned. That is, tokens largely attend to other tokens that occurred around the same window of time (Figure 2BC), an effect that is also visible when examining the raw attention matrices W_l (Appendix).

Although attention rollout allowed us to better understand the benefits of MultiMAE pretraining, it is unclear how input modality drop affects representations. To further investigate this, we next analyze individual embedding units in the model to see how tuned they are to different modalities. We use relative source variance (RSV), which quantifies the variance in the activity of a unit due to a particular input modality [Kleinman et al., 2023]. As an example, assume we want to calculate RSV due to EEG. First, let $x_{EEG} \sim X_{EEG}$ be a sample of EEG data from the dataset (with similar notation for all other modalities). The source variance of a unit a due to EEG when all other modalities j are fixed at samples x_j is defined as

$$SV_a(X_{EEG}, x_{EMG}, x_{EOG}, x_{ECG}) = \\ \text{Var}(f(X_{EEG}, X_{EMG} = x_{EMG}, X_{EOG} = x_{EOG}, X_{ECG} = x_{ECG})_a)$$

where f gives the output embedding from the encoder, averaged over tokens. Symmetrically, source variance can also be defined for the other modalities. Taking the softmax over these source variances for a unit a gives the relative source variance of a. Thus if unit a is uniformly tuned to all input sources, it would have a RSV value of 0.25 for each modality.

We first measure the RSV values of embedding units in the model trained from scratch on the sleep staging task (Figure 2D). We find that representations become more tuned for EEG in the later layers of the model. This is likely because EEG is more informative for the task and thus the decoder places greater emphasis on EEG over the other modalities. We next measure the RSV values for the MAE-pretrained model (Figure 2E). Interestingly, we see that across layers, units in the model become increasingly tuned to EMG input. This is likely because the model struggles most to reconstruct EMG (Appendix) and thus places greater representation weight onto that modality. In contrast, the model trained with MAE and modality drop is equally tuned to all modalities across all layers (Figure 2F).

5 Limitations and Discussion

We have shown the strength of a foundation model-style approach using physiological data with a diverse set of downstream tasks. We compare a variety of approaches and argue that explicitly incorporating objectives that promote cross-modal reconstruction greatly improves representation quality for solving downstream tasks. Specifically, we find that incorporating input modality drop is a simple, yet especially effective strategy. We note that making comparisons with other datasets would be additionally informative, especially since multimodal fusion strategies are often dependent on the dataset and task at hand [Ma et al., 2022]. In addition, developing a large range of downstream tasks will provide better insights into the strengths of different pretraining strategies and help identify those that are especially useful for general purpose training.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

Tarik Arici, Mehmet Saygin Seyfioglu, Tal Neiman, Yi Xu, Son Train, Trishul Chilimbi, Belinda Zeng, and Ismail Tutar. Mlim: Vision-and-language model pre-training with masked language and image modeling. *arXiv preprint arXiv:2109.12178*, 2021.

H Ceren Ates, Peter Q Nguyen, Laura Gonzalez-Macia, Eden Morales-Narváez, Firat Güder, James J Collins, and Can Dincer. End-to-end design of wearable sensors. *Nature Reviews Materials*, 7(11): 887–907, 2022.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

Joseph W Burns, Flavia B Consens, Roderick J Little, Karen J Angell, Sid Gilman, and Ronald D Chervin. Emg variance during polysomnography as an assessment for rem sleep behavior disorder. *Sleep*, 30(12):1771–1778, 2007.

Stefano Canali, Viola Schiaffonati, and Andrea Aliverti. Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness. *PLOS Digital Health*, 1(10):e0000104, 2022.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Federico Del Pup and Manfredo Atzori. Applications of self-supervised learning to biomedical signals: A survey. *IEEE Access*, 2023.

Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora Salim, and Akhil Mathur. Latent masking for multimodal self-supervised learning in health timeseries. *arXiv preprint arXiv:2307.16847*, 2023.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

LX Feng, X Li, HY Wang, WY Zheng, YQ Zhang, DR Gao, and MQ Wang. Automatic sleep staging algorithm based on time attention mechanism. front hum neurosci 15: 692054, 2021.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *arXiv preprint arXiv:2301.05712*, 2023.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. *arXiv preprint arXiv:2205.15465*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 378–386, 2020.

Vanessa Ibáñez, Josep Silva, and Omar Cauli. A survey on sleep assessment methods. *PeerJ*, 6: e4849, 2018.

Sheikh MA Iqbal, Imadeldin Mahgoub, E Du, Mary Ann Leavitt, and Waseem Asghar. Advances in healthcare wearable devices. *NPJ Flexible Electronics*, 5(1):9, 2021.

In Cheol Jeong, David Bychkov, and Peter C Searson. Wearable devices for precision medicine and health state monitoring. *IEEE Transactions on Biomedical Engineering*, 66(5):1242–1258, 2018.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods for multisensory integration in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24296–24305, 2023.

Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare. *arXiv preprint arXiv:2402.02460*, 2024.

Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

George Mathew, Daniel Barbosa, John Prince, and Subramaniam Venkatraman. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. 2024.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in Neural Information Processing Systems*, 32, 2019.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, 2022.

Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Contrastive pre-training for multimodal medical time series. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Santosh Kumar Satapathy, Biswajit Brahma, Baidyanath Panda, Paolo Barsocchi, and Akash Kumar Bhoi. Machine learning-empowered sleep staging classification using multi-modality signals. *BMC Medical Informatics and Decision Making*, 24(1):119, 2024.

Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, page 102040, 2023.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Y Zou. Sleepfm: Multi-modal representation learning for sleep across ecg, eeg and respiratory signals. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Shakti N Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. The evolution of multimodal model architectures. *arXiv preprint arXiv:2405.17927*, 2024.

Min Wu and Jake Luo. Wearable technology applications in healthcare: a literature review. *Online J. Nurs. Inform*, 23(3), 2019.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1528–1538, 2023.

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: Yes, the abstract and introduction match the claims made in the paper.

   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: Yes, limitations are explicitly mentioned in the discussion section. Throughout the results, we also discuss weaknesses in our approach and ways they can be improved.

   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: We do not have theoretical results in this paper.

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: Yes, details of the implementation needed to reproduce the paper are mentioned in the methods and the appendix.

   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: [Yes]

Justification: The data is an open source dataset that has been used across many works. The code will be made publicly available after the double-blind review process.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, these details are provided in the methods and appendix. Furthermore, publicly provided code will have all training details.

   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: Yes, error bars, number of samples, and the type of error bar statistic are all provided with each figure.

   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: Yes, the appendix provides thorough detail of compute resources used and the time experiments took.

   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: Yes, we have reviewed the code of ethics and our paper conforms with this code.

   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: Yes, we have a healthcare-centered motiviation which we discuss thoroughly in the introduction.

    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: [NA]

    Justification: The paper does not pose a risk in terms of data or model misuse. The data is already openly available, and the model is centered around interpreting this data.

    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: Yes, we give credit to existing data and code frameworks wherever relevant.

    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: [NA]

    Justification: We do not introduce new assets.

    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: [NA]

    Justification: We did not collect new data with human subjects.

    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: We did not collect new data with human subjects.

    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.