RespiraMFM: A Multimodal Foundation Model with Contrastive Audio-Language Alignment for Respiratory Disease Identification

Anonymous ACL submission

Abstract

Respiratory diseases remain a leading cause of global mortality, where timely and accurate diagnosis is critical to improving patient outcomes and reducing healthcare burdens. While prior work has explored audio-based models for respiratory disease detection, such unimodal approaches often suffer from limited generalizability and diagnostic precision. In this paper, we propose `RespiraMFM`, a Multimodal Foundation Model that integrates respiratory sounds with patient medical history and symptoms to enhance diagnostic accuracy and disease detection capabilities. We introduce an effective contrastive alignment strategy for audio-text multimodal integration, allowing the model to learn better cross-modal representations between respiratory sounds and corresponding textual clinical information. We evaluate `RespiraMFM` across five major respiratory diseases using seven real-world datasets in both supervised fine-tuning and zero-shot settings, achieving a 9.72% improvement in AUROC on supervised tasks and a 20.36% gain on zero-shot tasks over existing baselines. These findings underscore the potential of our framework to advance early diagnosis and improve clinical decision-making in respiratory disease management.

1 Introduction

Respiratory diseases, such as COVID-19, tuberculosis (TB), chronic obstructive pulmonary disease (COPD), asthma, and pneumonia, remain a leading cause of morbidity and mortality worldwide (Weinberger et al., 2020). Most existing works (Baur et al., 2024; Zhang et al., 2024a) on detecting those respiratory diseases rely solely on audio inputs such as coughing sounds or stethoscope recordings. However, their performance is often constrained by the limited information that audio data alone can provide.

To mitigate the constraints of relying solely on audio inputs, researchers have proposed multimodal methods (Kim et al., 2024; Zhang et al., 2024b) that combine respiratory audio with relevant clinical information, such as symptoms (e.g., fever, fatigue, chest pain) and lifestyle factors such as smoking history. For instance, BTS (Kim et al., 2024) concatenates learned representations from audio and text encoders for respiratory disease identification. RespLLM (Zhang et al., 2024b) utilizes a large language model (LLM) as the text encoder alongside a separate audio encoder, with a linear projector for matching dimensions.

Although combining audio input with text in a multi-modal setting has been widely adopted in recent works (Ma et al., 2024; Zhang et al., 2025), directly applying the existing approach in the respiratory disease detection task poses a unique challenge. Prior studies on audio–text alignment have primarily focused on spoken language, where audio carries rich semantic and syntactic information that can be naturally aligned with text. In contrast, the respiratory disease identification task involves fusing non-linguistic acoustic biomarkers (i.e., coughs, wheezes, and crackles) with free-form patient symptoms text. Existing work on respiratory LLMs, such as RespLLM (Zhang et al., 2024b), often adopts the same straightforward fusion techniques used for spoken audio, typically relying on simple feature concatenation or a trainable linear projector within a unified end-to-end framework. This strategy suffers from two critical limitations: First, because non-linguistic acoustic biomarkers like cough sounds are fundamentally misaligned with the text description of a patient's condition, the unified training objective struggles to converge on a stable shared representation, resulting in failure to leverage both modalities effectively. Second, due to this inefficient fusion mechanism being tightly coupled to the training data, these models only perform well on in-domain diseases and fail in zero-shot scenarios when encountering

new, unseen diseases.

To address these challenges, we introduce RespiraMFM, a two-stage, decoupled training architecture for respiratory disease identification. In our two-stage approach, the first stage is dedicated to Modality Alignment. A lightweight linear projector is contrastively trained to learn a robust, non-linear mapping that projects the high-dimensional audio features directly into the semantic embedding space of the LLM. Reflecting the success of contrastive methods in improving zero-shot performance (Wu et al., 2023), this pre-training provides better initialization for the weights of the projector and ensures the acoustic features are semantically anchored to the correct symptom concepts. For the second stage, we freeze the learned projector module, and the resulting aligned embeddings are then passed to the LLM to generate the final disease prediction.

We evaluate RespiraMFM using seven real-world datasets that cover five of the most common respiratory diseases: COVID-19, TB, COPD, asthma, and pneumonia. We highlight four of our findings: (1) RespiraMFM consistently outperforms state-of-the-art multimodal baselines on respiratory disease identification by achieving a 9.72% improvement in AUROC on supervised tasks and a 20.36% improvement on zero-shot tasks. (2) RespiraMFM achieves superior generalization capabilities and effectively detects unseen respiratory diseases without requiring any training samples of those diseases. (3) RespiraMFM significantly reduces the training data requirement, achieving comparable performance with an order of magnitude less training data compared to the baselines. (4) Our contrastive alignment module effectively unifies audio and text modalities, leading to consistent AUROC improvements across all tasks compared to models without this module.

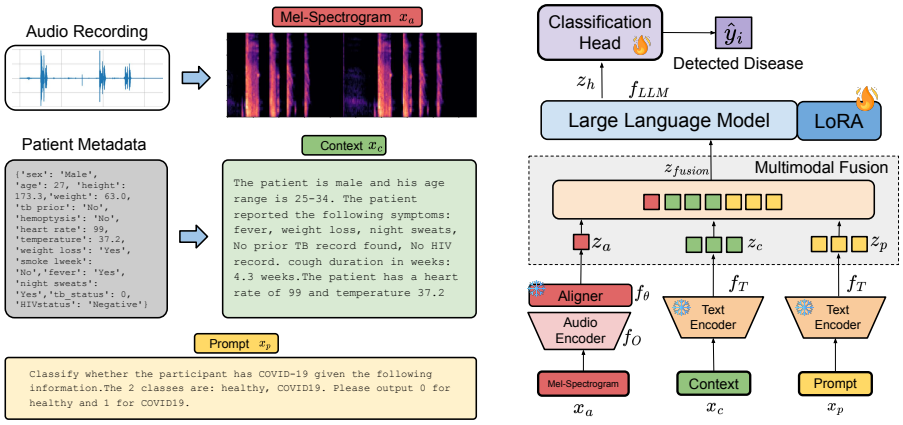
2 Related Work

2.1 Single-Modal Models

The majority of existing respiratory disease identification methods (Yang et al., 2020; Ma et al., 2020; Chang et al., 2022) rely solely on audio inputs such as cough sounds or stethoscope recordings. Bae et al. (2023) introduce a contrastive learning framework to enhance respiratory sound classification using Audio Spectrogram Transformer (AST). By mixing spectrogram patches generated from raw audio data and applying contrastive loss, the model learns robust and discriminative features, which are subsequently passed to a linear classifier for respiratory disease identification. OPERA (Zhang et al., 2024a) curates large-scale unlabeled respiratory audio datasets and pretrains three foundational models using self-supervised learning. Among them, OPERA-CT, the best-performing model, is a contrastive learning–based transformer model, which is used as general-purpose feature extractors for respiratory disease classification tasks. HeAR (Health Acoustic Representations) (Baur et al., 2024) introduces a self-supervised generative learning-based framework trained on a large corpus of health-related audio data. By leveraging generative objectives during pretraining, HeAR learns generalizable audio representations which is utilized for downstream disease diagnosis tasks via simple linear probes. Despite the promising results of single-modal models, their performance is limited by the information available from audio data alone.

2.2 Multimodal Models

Unlike single-modal models, multimodal models combine audio data with textual information such as patient symptoms and medical history, leading to more accurate diagnoses. BTS (Kim et al., 2024) introduces a text-audio model that combines respiratory sounds with metadata transformed into descriptive text. It uses the Contrastive Language-Audio Pretraining (CLAP) (Elizalde et al., 2023) model to extract features from both modalities, followed by a linear classifier for respiratory disease classification. However, the use of a basic linear classifier limits its ability to generalize in zero-shot scenarios or when encountering new diseases. To date, RespLLM (Zhang et al., 2024b) is one of the early efforts that applies a multimodal LLM framework integrating text and audio representations for respiratory disease prediction. Their approach utilizes a pretrained encoder to extract audio and text features and a trainable linear projector to align the feature dimensions with LLM. However, since each modality encoder is trained separately, the resulting representations are often distinct and may not be directly compatible across modalities. While a linear projector can align the encoder output dimensions with those expected by the LLM, it does not ensure semantic alignment between modalities. To address these limitations, we propose a contrastive alignment module that facilitates more effective integration by aligning audio and text representations in a shared semantic space. Our approach goes



(a) Data Curation

(b) Multimodal foundation model for respiratory disease identification.

Figure 1: Overview of RespiraMFM.

beyond mere dimensional alignment, aiming to establish a shared representation space that enables effective integration of multimodal information.

3 RespiraMFM

3.1 Overview

Figure 1 provides an overview of the proposed `RespiraMFM` framework. In the data curation stage shown in Figure 1(a), given the multimodal respiratory datasets, we extract and pre-process the raw audio data and the corresponding patient symptoms to construct the instruction tuning data for respiratory disease identification. As shown in Figure 1(b), the curated multimodal data are first processed by the audio and text encoders to get audio and text representations, respectively. One key component of our framework is the alignment module that reduces the domain mismatches between audio features and the language model embeddings. The alignment module is trained separately via contrastive learning. Upon completion of training, the alignment module is frozen and incorporated into the instruction tuning stage. During instruction tuning, the curated data are passed through each encoder to obtain modality-specific representations, which are then fused by concatenating them. The resulting multimodal representation is subsequently fed into the LLM to generate predictions for respiratory disease classification.

3.2 Data Curation

The multimodal respiratory datasets consist of both respiratory audio recordings and the corresponding patient-reported symptoms in either JSON or tabular format. The objective of data curation is to create the instruction tuning data by pre-processing

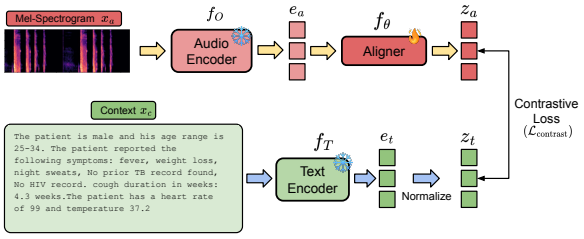


Figure 2: Illustration of contrastive learning-based audio-text alignment.

the respiratory audio recordings, generating instruction prompts, and converting patient-reported symptoms into structured textual representations. For audio recordings, each recording was normalized to 8 seconds in length by either truncating longer recordings or padding shorter ones through repetition. The audio signals are then processed with a 64ms Hann window with a 32ms step size, and subsequently converted into mel spectrograms denoted as x_a using the features extracted from a pre-trained OPERA-CT encoder (Zhang et al., 2024a). Patient metadata varies across datasets in terms of structure and format. As shown in Figure 1(a), we select relevant symptoms (Table 5) from the tabular data and apply a standardized template to generate a textual representation x_c. We utilize task-specific prompts x_p such as - *"Classify whether the participant has COVID-19 given the following information.The 2 classes are: healthy, COVID19. Please output 0 for healthy and 1 for COVID19"* to guide the LLM in producing disease classification outputs.

3.3 Contrastive Learning-based Audio-Text Aligning

We introduce a contrastive learning-based audio-text alignment module to align audio features with language model embeddings effectively. Specifically, we employ a pre-trained audio encoder that generates 768-dimensional embeddings from the input audio. In contrast, large language models typically operate with higher-dimensional input embeddings. Therefore, it is necessary to match the audio dimension to use as input into the LLM. Prior work (Zhang et al., 2024b) addresses this by introducing a trainable linear projector to map audio embeddings into the higher-dimensional space required by the LLM. However, given the fundamental differences between the audio and text encoders in both architecture and representational semantics, simple dimensional alignment may be insufficient to achieve effective multimodal fusion (Lyu et al., 2023, 2024). To enable more effective cross-modal alignment between audio and text, we adopt a contrastive learning strategy—an approach shown to yield powerful multimodal representations in models like CLIP (Radford et al., 2021). As shown in Figure 2, we compute text embeddings $\mathbf{e}_t \in \mathbb{R}^d$ using a frozen LLM, where $\mathbf{e}_t = f_T(x_c)$, f_T is the text encoder (LLM), x_c is the textual context, and d is the embedding dimension of LLM. Similarly, audio embeddings $\mathbf{e}_a \in \mathbb{R}^{768}$ are extracted using a frozen OPERA encoder, where $\mathbf{e}_a = f_O(x_a)$, f_O is the pre-trained Opera-CT audio encoder, and x_a is the mel-spectrogram of the raw audio data.

A lightweight projection head $f_\theta : \mathbb{R}^{768} \to \mathbb{R}^d$ is trained to map audio embeddings into the same semantic space as the text embeddings. The training objective minimizes a contrastive loss that encourages matched audio-text pairs to be close while pushing unmatched pairs apart.

Formally, for a batch of N paired samples, we define the normalized embeddings as:

$$\mathbf{z}_i^a = \frac{f_\theta(\mathbf{e}_i^a)}{\|f_\theta(\mathbf{e}_i^a)\|}, \quad \mathbf{z}_i^t = \frac{\mathbf{e}_i^t}{\|\mathbf{e}_i^t\|}.$$

The contrastive loss (Chen et al., 2020) is given by:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathbf{z}_i^a \cdot \mathbf{z}_i^t/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z}_i^a \cdot \mathbf{z}_j^t/\tau)},$$

where τ is a temperature scaling factor. By training this projection head with contrastive supervision, we achieve better semantic alignment across modalities while keeping the audio and text encoders frozen. The model architecture and additional training details about the aligner module are presented in Appendix D.

3.4 Instruction Tuning

We employ instruction tuning to guide the LLM in understanding and following task-specific prompts that connect the multi-modal input and the corresponding diagnostic outcomes. The core components of instruction tuning are described below.

Multimodal Fusion: The audio and text features are fused at the embedding level by concatenation. During this stage, we utilize the contrastively trained alignment module (f_θ) from the previous step (§3.3), keeping its weights frozen to preserve the learned representations. Similarly, the text encoder (f_T) and the audio encoder (f_O) are also kept frozen during this stage. The inputs include mel-spectrogram (x_a), curated patient symptom descriptions as contextual information (x_c), and a task-specific prompt (x_p). Audio embeddings $z_a \in \mathbb{R}^d$ are extracted via the audio encoder and subsequently projected to match the input dimensionality of the LLM.

$$z_a = f_\theta(f_O(x_a))$$

Simultaneously, the prompt and contextual text are processed through the LLM's encoder to obtain their respective representations, denoted as $z_p \in \mathbb{R}^d$ and $z_c \in \mathbb{R}^d$, corresponding to the prompt and context embeddings.

$$z_p = f_T(x_p), \quad z_c = f_T(x_c).$$

Finally, we concatenate the audio (z_a), prompt (z_p), and context (z_c) embeddings to get a combined embedding of a longer sequence:

$$z_{fusion} = z_a \parallel z_p \parallel z_c$$

where $\parallel$ denotes concatenation operation and $z_{fusion} \in \mathbb{R}^d$.

Large Language Model: We utilize Phi-2[1], a 2.7B parameter model, as the backbone LLM. To adapt it for our classification task, we extend the model by appending a linear classification head atop the transformer architecture. We first form the multimodal fusion embeddings by concatenating audio and text representations. These fused embeddings are then fed into the LLM to produce a sequence of hidden states. A pooling layer is then applied

[1]https://huggingface.co/microsoft/phi-2

Table 1: Summary of the datasets and tasks.

Task ID	Dataset	Disease	#Train/Test
T1	UK COVID-19 (Coppock et al., 2024)	COVID-19	20717/11121
T2	Coughvid (Orlandic et al., 2021)	COVID-19	7958/2464
T3	TBscreen (Sharma et al., 2024)	TB	20302/8051
T4	ICBHI (Rocha et al., 2019)	COPD	462/366
T5	Coswara (Bhattacharya et al., 2023)	COVID-19	-/1747
T6	CodaTB (Huddart et al., 2024)	TB	-/2053
T7	KAUH (Fraiwan et al., 2022)	COPD	-/132
T8	KAUH (Fraiwan et al., 2022)	Asthma	-/201
T9	KAUH (Fraiwan et al., 2022)	Pneumonia	-/120

to obtain the latent representation z_h. Specifically, we adopt the default pooling strategy used in `AutoModelForSequenceClassification`, which selects the hidden state corresponding to the final token in the sequence. Finally, a linear classification head is applied to the pooled representation to produce prediction scores for different respiratory disease identification tasks.

$$z_h = Pool_{final}(f_{LLM}(z_{fusion}))$$

This vector z_h is then passed through a classification head comprising fully connected layers, followed by a softmax function to produce class probability distributions. The model is trained using cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where y_i and $\hat{y}_i$ are the true and predicted probabilities for class i, respectively.

Training Details: The instruction tuning process combines task-specific instructions x_p with multimodal audio (x_a) and text (x_c) inputs to ensure the model generates outputs that align with the desired response format. Additionally, we use LoRA (Low-Rank Adaptation) (Hu et al., 2021), a parameter-efficient fine-tuning (PEFT) technique designed to preserve the inherent knowledge of a pre-trained LLM. The model was trained for 40 epochs, and the training configuration further optimizes LoRA with parameters like a rank (r) of 16, scaling factor (α) of 32, and a dropout of 0.1.

4 Experimental Setup

4.1 Datasets and Tasks

We evaluate the performance of `RespiraMFM` using seven real-world datasets, covering five of the most common respiratory diseases: COVID-19, TB, COPD, asthma, and pneumonia. These datasets include both respiratory audio recordings (e.g., coughing sound, stethoscope sound) and the associated metadata, such as patient-reported symptoms and medical history. Based on these datasets, we construct nine respiratory disease identification tasks as summarized in Table 1. Datasets associated with tasks T1 through T4 are used for training and in-domain evaluation using held-out test sets, while datasets associated with tasks T5 through T9 are reserved for zero-shot evaluation. For each task, the model is trained on the combined training data from T1 to T4. For example, in T5, the model is trained using all the training sets from tasks T1 to T4 and evaluated on the T5 test set. Notably, T8 and T9 involve entirely new diseases (asthma and pneumonia) not seen during training, allowing us to assess the model's generalization ability to previously unseen conditions in a zero-shot setting. Details of each dataset and task are provided in Appendix A.

4.2 Baselines and Evaluation Metrics

Baselines: We compare RespiraMFM against three state-of-the-art multimodal baselines: Qwen-2 Audio (Chu et al., 2024), BTS (Kim et al., 2024) and RespLLM (Zhang et al., 2024b). More details about the baselines are included in Appendix C.

Evaluation Metrics: To ensure fair comparison, we follow prior works on respiratory disease detection to use the Area Under the Receiver Operating Characteristic Curve (AUROC) (Janssens and Martens, 2020) as the evaluation metric for all the tasks. To ensure robust evaluation, each result was obtained through three independent runs. The mean and standard deviation of the AUROC scores across these runs are reported.

4.3 Implementation Details

We utilized PyTorch 2.3.0, transformers 4.47.1 (Wolf et al., 2020), and accelerated on four NVIDIA A100-80G GPUs. The training process uses a batch size of 16.

5 Results

5.1 Overall Performance

First, we compare the performance of `RespiraMFM` with the baselines under the supervised learning setting on the held-out test

Table 2: AUROC comparison for respiratory disease recognition task. Results are shown in $mean \pm std$ format of three individual runs. The teal color indicates the best results. The values in parentheses represent the relative improvement (%) of RespiraMFM over the strongest baseline for each task.

Task ID	Dataset	Disease	Qwen-2 Audio	BTS	RespLLM	RespiraMFM (ours)
T1	UK COVID-19	COVID-19	0.855 ± 0.018	0.898 ± 0.010	0.881 ± 0.005	0.906 ± 0.002 (↑ 0.89 %)
T2	Coughvid	COVID-19	0.561 ± 0.009	0.595 ± 0.014	0.613 ± 0.011	0.648 ± 0.011 (↑ 5.7 %)
T3	TBscreen	TB	0.334 ± 0.043	0.568 ± 0.019	0.687 ± 0.016	0.817 ± 0.014 (↑ 18.9 %)
T4	ICBHI	COPD	0.614 ± 0.005	0.880 ± 0.004	0.833 ± 0.007	0.998 ± 0.000 (↑ 13.4 %)

Table 3: AUROC comparison for the respiratory disease recognition task of zero-shot prediction on new dataset. Results are shown in $mean \pm std$ format of three individual runs. The teal color indicates the best results. The values in parentheses represent the relative improvement (%) of RespiraMFM over the strongest baseline for each task.

Task ID	Dataset	Task	Qwen-2 Audio	BTS	RespLLM	Zero-shot RespiraMFM (ours)
T5	Coswara	Covid	0.813 ± 0.035	0.901 ± 0.008	0.90 ± 0.006	0.904 ± 0.005 (↑ 0.33 %)
T6	CodaTB	TB	0.527 ± 0.012	0.645 ± 0.016	0.669 ± 0.019	0.670 ± 0.012 (↑ 0.15 %)
T7	KAUH	COPD	0.581 ± 0.013	0.491 ± 0.014	0.425 ± 0.011	0.768 ± 0.005 (↑ 32.19 %)
T8	KAUH	Asthma	0.458 ± 0.010	0.418 ± 0.016	0.399 ± 0.010	0.617 ± 0.014 (↑ 34.72 %)
T9	KAUH	pneumonia	0.301 ± 0.041	0.595 ± 0.020	0.400 ± 0.021	0.800 ± 0.013 (↑ 34.45 %)

sets of the training datasets on tasks T1 through T4. The results are summarized in Table 2. As shown, RespiraMFM consistently outperforms all other baselines across all four tasks. Overall, the average AUROC RespiraMFM has achieved over tasks T1 through T4 is 0.842, representing 42.4% improvement over Qwen2-Audio (average AUROC: 0.591), 16.4% improvement over BTS (average AUROC: 0.723) and 10.4% improvement (average AUROC: 0.762) over RespLLM. These results demonstrate the strong performance of RespiraMFM in identifying a wide range of respiratory diseases, advancing the state of the art.

5.2 Zero-Shot Performance

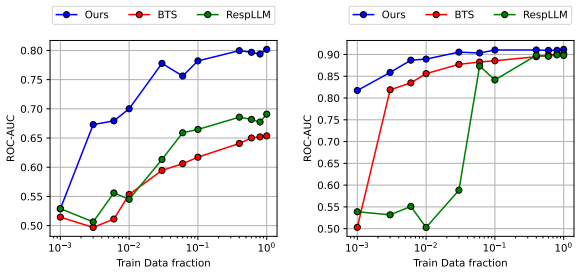
Next, we evaluate the zero-shot performance of RespiraMFM under the following two scenarios.

Unseen Datasets: Regarding the unseen datasets condition, we compare RespiraMFM with other multi-modal baselines in performing tasks T5, T6 and T7. In these tasks, the datasets used for evaluation are not seen during training, though the target diseases remain the same. Specifically, the models are trained on one or more datasets for a given disease and evaluated on a different, unseen dataset for the same condition. For example, in task T5, the training data includes other COVID-19 datasets such as UKCOVID-19 and CoughVid, and is evaluated on the unseen Coswara dataset. As shown in Table 3, the proposed RespiraMFM consistently outperforms all other multi-modal baselines on these unseen datasets. Specifically, our average AUROC over these tasks is 0.75, outperforming Qwen2-Audio by 38.8% (average 0.54), BTS by 22.9% (average 0.61) and RespLLM by 33.9% (average 0.56) on average AUROC.

Unobserved Respiratory Diseases: Regarding the unobserved respiratory diseases, we further compare RespiraMFM with BTS and RespLLM on the prediction of asthma (T8) and pneumonia (T9). In both tasks, the models are trained on datasets from T1 to T4, none of which include instances of asthma or pneumonia. As shown in Table 3, despite having no disease-specific training data for these conditions, RespiraMFM consistently outperforms all the baselines. Specifically, RespiraMFM achieves a 34.72% and 34.45% relative improvement in asthma and pneumonia prediction over the other baselines, respectively. Overall, these results suggest that RespiraMFM generalizes effectively across datasets and to previously unseen respiratory diseases.

5.3 Effects of Data Scaling

To assess how the training dataset size impacts the model performance, we conducted experiments on Task T1 by systematically varying the number of training examples. In this experiment, the model was trained on a combined set of UKCOVID-19, Coughvid, TBscreen, and ICBHI datasets and evaluated on the held-out test of the UKCOVID-19 dataset. Starting with a full training set of 49,439



(a) Single-Modal (b) Multimodal

Figure 3: Effect of dataset scaling.

	Mild or No symptoms	Moderate symptoms	Healthy	Total
Audio	0.3576	0.3571	0.7266	0.6102
Text	0.3294	0.619	0.9766	0.7934
Audio+Text	**0.4047**	**0.6587**	**0.9849**	**0.8203**

Table 4: Performance comparison of audio-only, text-only, and multimodal (audio+text) models across different patient groups in the Coswara dataset. **Bold** indicates the best performance and underlined indicates the second-best.

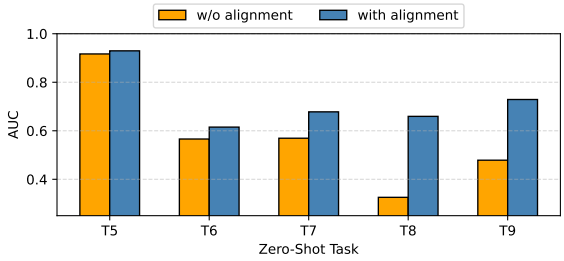


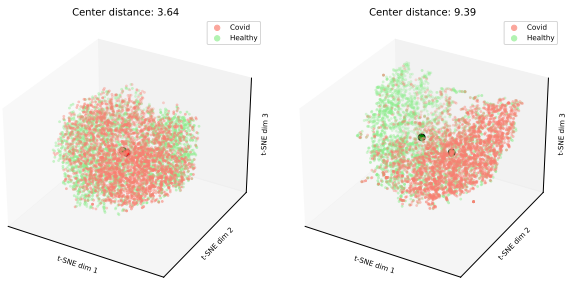
Figure 4: AUC of zero-shot disease detection task with and without alignment module

samples, we randomly sampled subsets at varying fractions and compared our model with the baselines on the same test set. We explored two configurations for this experiment: a single-modal setup using only audio features as input, and a multi-modal setup that integrates both audio and textual features as input. The results are shown in Figure 3. Figure 3a, which corresponds to the single-modal setting using only audio input, shows a clear trend of improved performance with increasing training samples, indicating that larger datasets lead to better performance. Our model consistently outperforms both BTS and RespLLM across all data fractions, with notably strong performance even at low data availability. While all models benefit from more data, ours maintains a consistent lead. In contrast, Figure 3b illustrates the multi-modal configuration, where both audio and text features are used as input. Here, our model rapidly approaches peak performance with minimal training data and significantly outperforms the baselines across nearly all data scales. These results highlight the strength of multi-modal integration, especially in clinical contexts where labeled data is often limited. The findings suggest that multi-modal models are particularly well-suited for deployment in resource-constrained healthcare settings, offering high diagnostic performance even with sparse training data.

5.4 Ablation Study

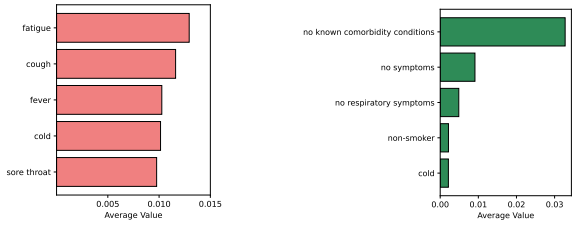
Uni-Modality vs. Multi-Modality: To assess the effectiveness of multimodal integration compared to unimodal inputs, we conducted experiments on Task T5, aiming to understand whether combining audio and textual information offers complementary benefits that improve diagnostic performance beyond what a single modality can achieve alone. In this experiment, the model is trained on the combined data from all available training datasets and evaluated in a zero-shot setting on the Coswara dataset. We select the Coswara dataset for this experiment because it provides both disease labels and additional metadata describing patient health status, including severity levels such as asymptomatic (no symptoms), mild, moderate, and healthy. We group these into three broad categories—mild or no symptoms, moderate symptoms, and healthy—and evaluate models in three configurations: audio-only input (uni-modal), text-only input (uni-modal), and multimodal input combining both audio and text. Accuracy is used as the evaluation metric for all configurations. As shown in Table 4, for cases with mild or no symptoms, the audio-only model outperforms the text-only model based on the symptom information. Conversely, the text-only model performs better compared to the audio-only model for symptomatic and healthy individuals. On the other hand, the multimodal model, which integrates both audio and text information, consistently outperforms both unimodal models across all severity levels and on the overall dataset. In summary, these results demonstrate the clear advantage of combining multiple modalities for improved disease prediction.

Analysis on alignment module: To evaluate the effectiveness of the contrastive alignment mod-



(a) without alignment (b) with alignment

Figure 5: t-SNE visualization of audio embeddings for ukcovid-19 dataset. The left panel shows audio embeddings obtained without alignment, while the right panel shows those with alignment. The green points represent sample from healthy patients and red points represent sample from covid patients.



(a) Covid Patients (b) Healthy Patients

Figure 6: Average attention weight of top-5 tokens for both healthy and covid-19 patients on coswara dataset.

ule introduced in §3.3, we conducted experiments across all the zero-shot tasks (T5-T9) by training the model both with and without this component. For the baseline setting (without alignment), we used a standard linear projector, as commonly applied in prior work (Zhang et al., 2024b), and trained it jointly with the LLM in a single-stage fine-tuning process. In contrast, for the alignment setting, we first trained the projector using a contrastive loss and then kept it frozen during LLM fine-tuning. The results are summarized in Figure 4. As shown, the alignment module consistently improves AUC scores across all tasks, suggesting that it provides more informative feature representations for instruction tuning. To further understand why the improvement on zero-shot tasks occurs, we analyzed the learned audio embeddings from the projectors and visualized them using 3D t-SNE (Figure 5). The left panel (5a) shows embeddings obtained without alignment, while the right panel (5b) shows those with alignment. Samples from COVID patients are colored in red, and healthy samples are colored in green. The embeddings with alignment form clearer clusters and are more separable between the two classes compared to the baseline. This enhanced separability allows the model to learn more discriminative and semantically aligned audio features, thereby improving generalization and leading to better zero-shot performance.

Interpretability analysis: To evaluate the interpretability of our models, we extracted the `[CLS]` token attention and measured the average attention weight assigned to each context token. As shown in Figure 6, we conducted experiments separately for healthy and COVID-positive patients and report the average attention across their respective tokens or groups of tokens. From the plot, we can see that for COVID-positive patients, most of the attention is concentrated on tokens representing fever, cough, fatigue, and other symptomatic indicators, which is clinically expected. On the other hand, for healthy patients, most of the attention is placed on token groups that explicitly represent the absence of symptoms (e.g., no symptoms, no respiratory symptoms, non-smoker). This indicates that the model not only attends to disease-related symptoms when present but also leverages the absence of symptoms for decision-making. These observations highlight that the model's focus aligns with medically relevant patterns, indicating a degree of interpretability in an otherwise black-box architecture.

6 Conclusion

In this paper, we introduced RespiraMFM, a multimodal foundation model designed to detect respiratory diseases by integrating respiratory sound recordings with patient-reported symptoms and medical history. We also proposed an effective method for multimodal alignment of text and audio input, demonstrating strong performance across nine tasks involving five major respiratory diseases using diverse real-world datasets. We also showed that the model can maintain high diagnostic accuracy even with limited training data, making it suitable for deployment in data-scarce healthcare environments. Overall, RespiraMFM offers a scalable, non-invasive, and clinically relevant solution for early and accurate respiratory disease detection, with the potential to support medical professionals and improve decision-making across a variety of healthcare settings.

7 Limitation

While our proposed multimodal foundation model shows strong performance across various respiratory disease detection tasks, it has some limitations. The model's effectiveness depends on the quality and consistency of symptom metadata, which can differ significantly between datasets and clinical environments. For instance, in Task T2, the model performs relatively lower compared to other COVID-19 detection tasks (T1 and T5), likely due to the limited or less informative symptom data available in the Coswara dataset, making accurate diagnosis more challenging. Additionally, although the model integrates audio and symptom data, incorporating additional modalities such as medical imaging or wearable sensor data could further improve its diagnostic accuracy and robustness.

8 Ethics Statement

We foresee no ethical concerns with our work. All the datasets used in this study were anonymized and excluded any participant identity information.

References

Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Se-Young Yun. 2023. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. *arXiv preprint arXiv:2305.14032*.

Sebastien Baur, Zaid Nabulsi, Wei-Hung Weng, Jake Garrison, Louis Blankemeier, Sam Fishman, Christina Chen, Sujay Kakarmath, Minyoi Maimbolwa, Nsala Sanjase, et al. 2024. Hear–health acoustic representations. *arXiv preprint arXiv:2403.02522*.

Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C Chandrakiran, Sahiti Nori, KK Suhail, Sadhana Gonuguntla, et al. 2023. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific Data*, 10(1):397.

Yi Chang, Zhao Ren, Thanh Tam Nguyen, Wolfgang Nejdl, and Björn W Schuller. 2022. Example-based explanations with adversarial attacks for respiratory sound analysis. *arXiv preprint arXiv:2203.16141*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Harry Coppock, George Nicholson, Ivan Kiskin, Vasiliki Koutra, Kieran Baker, Jobie Budd, Richard Payne, Emma Karoune, David Hurley, Alexander Titcomb, et al. 2024. Audio-based ai classifiers show no evidence of improved covid-19 screening over simple symptoms checkers. *Nature Machine Intelligence*, 6(2):229–242.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mohammad Fraiwan, Luay Fraiwan, Mohanad Alkhodari, and Omnia Hassanin. 2022. Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sophie Huddart, Vijay Yadav, Solveig K Sieberts, Larson Omberg, Mihaja Raberahona, Rivo Rakotoarivelo, Issa N Lyimo, Omar Lweno, Devasahayam J Christopher, Nguyen Viet Nhung, et al. 2024. a dataset of solicited cough sound for tuberculosis triage testing. *Scientific Data*, 11(1):1149.

A Cecile JW Janssens and Forike K Martens. 2020. Reflection on modern methods: Revisiting the area under the roc curve. *International journal of epidemiology*, 49(4):1397–1403.

June-Woo Kim, Miika Toikkanen, Yera Choi, Seoung-Eun Moon, and Ho-Young Jung. 2024. Bts: Bridging text and sound modalities for metadata-aided respiratory sound classification. *arXiv preprint arXiv:2406.06786*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. 2024. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26752–26762.

Yi Ma, Xinzi Xu, and Yongfu Li. 2020. Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation. In *Interspeech*, pages 2902–2906.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

Lara Orlandic, Tomas Teijeiro, and David Atienza. 2021. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156.

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

Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. 2019. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001.

Manuja Sharma, Videlis Nduba, Lilian N Njagi, Wilfred Murithi, Zipporah Mwongera, Thomas R Hawn, Shwetak N Patel, and David J Horne. 2024. Tbscreen: A passive cough classifier for tuberculosis screening with a controlled dataset. *Science Advances*, 10(1):eadi0282.

Daniel M Weinberger, Jenny Chen, Ted Cohen, Forrest W Crawford, Farzad Mostashari, Don Olson, Virginia E Pitzer, Nicholas G Reich, Marcus Russi, Lone Simonsen, et al. 2020. Estimation of excess deaths associated with the covid-19 pandemic in the united states, march to may 2020. *JAMA internal medicine*, 180(10):1336–1344.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zijiang Yang, Shuo Liu, Meishu Song, Emilia Parada-Cabaleiro, and Björn W Schuller. 2020. Adventitious respiratory classification using attentive residual neural networks.

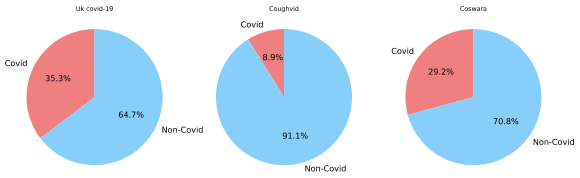
Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, and Haizhou Li. 2025. Soundwave: Less is more for speech-text alignment in llms. *arXiv preprint arXiv:2502.12900*.

Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. 2024a. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. *arXiv preprint arXiv:2406.16148*.

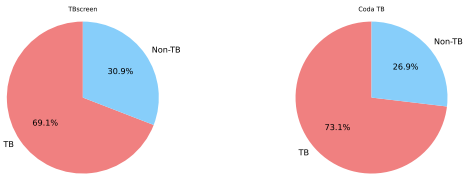
Yuwei Zhang, Tong Xia, Aaqib Saeed, and Cecilia Mascolo. 2024b. Respllm: Unifying audio and text with multimodal llms for generalized respiratory health prediction. *arXiv preprint arXiv:2410.05361*.

A Additional Details on Datasets

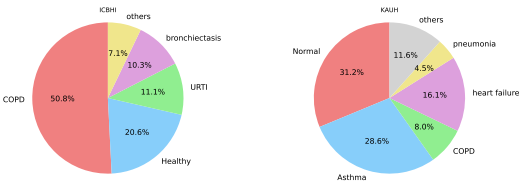
In this study, we used the following datasets: **UK COVID-19**: The UK COVID-19 Vocal Audio Dataset (Coppock et al., 2024) represents the largest collection of SARS-CoV-2 PCR-referenced audio recordings to date, compiled in the United Kingdom. The dataset features PCR test results linked to 70,794 out of 72,999 participants, with 24,155 of the 25,776 confirmed positive cases accurately documented. Notably, respiratory symptoms were reported by 45.62% of the participants, providing valuable symptomatic metadata for analysis. All the audio recordings were captured in the .wav format. In our study, we adopt the official train-test split released with the dataset.



(a) Class Distribution in covid datasets (UK covid-19, coughvid and coswara)



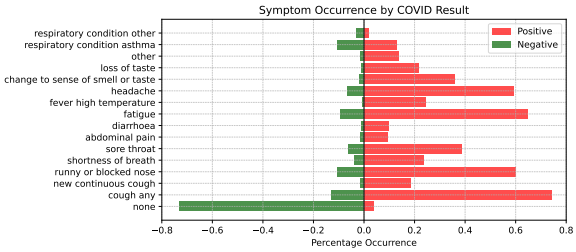
(b) Class Distribution in TB datasets (TBscreen and Coda TB)



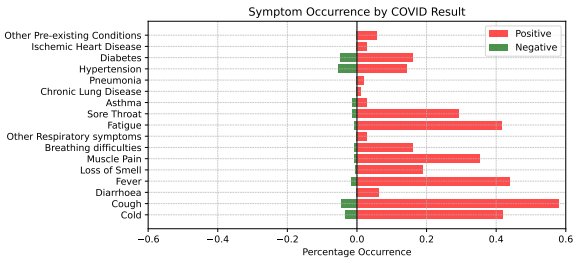
(c) Class Distribution in ICBHI and KAUH datasets

Figure 7: Class Distribution Across Datasets

Coswara: The Coswara dataset (Bhattacharya et al., 2023) is a diverse collection of respiratory sounds and detailed metadata, recorded between April 2020 and February 2022 from 2,635 individuals, including 1,819 SARS-CoV-2 negative, 674 positive, and 142 recovered cases. It features nine categories of respiratory sounds, covering variations of breathing, coughing, and speech, providing a rich dataset for analyzing respiratory health. In addition to audio recordings, the dataset includes comprehensive metadata, capturing demographic details such as age, gender, and geographic location, along with health-related information like symptoms, pre-existing respiratory conditions, comorbidities, and COVID-19 test status. We follow the official data split, which contains 70% samples for training, 15% for validation, and 15% for testing.



(a) UK-covid19 dataset



(b) Coswara dataset

Figure 8: Symptom Occurrence Distribution by COVID-19 Test Result in UK COVID-19 and Coswara Datasets.

COUGHVID: The COUGHVID dataset (Orlandic et al., 2021) is a large-scale, publicly available collection of over 25,000 crowdsourced cough recordings, covering a diverse range of ages, genders, geographic locations, and COVID-19 statuses. The database contains approximately 35 hours of audio recordings, comprising around 37,000 segmented cough samples. An automatic cough classifier was used to filter recordings, retaining only those with a minimum probability of 0.8 of containing cough sounds. The final distribution of labeled recordings was as follows: 25% COVID-positive cases, 35% symptomatic cases, 25% healthy individuals, and 15% with no reported health status.

TBscreen: The TBscreen dataset (Sharma et al., 2024) was collected in Nairobi and comprises cough recordings from 149 subjects diagnosed with pulmonary tuberculosis (TB) and 46 control subjects with other respiratory illnesses. The dataset includes a total of 33,000 passive coughs and 1,600 forced coughs, all recorded in a controlled setting to ensure consistency across subjects with similar demographics. To standardize the data for applica-

tions, each cough recording was processed to have a fixed duration of one second. Longer recordings were segmented into multiple one-second audio files, while shorter recordings were centered and padded with zeros to maintain uniformity.

CodaTB: The CodaTB dataset (Huddart et al., 2024) is a large, multi-country collection of cough sounds from individuals undergoing evaluation for tuberculosis (TB). It comprises over 700,000 cough recordings from 2,143 participants, along with detailed demographic, clinical, and microbiological diagnostic information. The dataset was collected as part of broader TB research studies, where participants underwent a baseline questionnaire, clinical examination, and sputum collection for TB testing at the time of enrollment. Comprehensive metadata accompanies the cough recordings, including age, gender, height, weight, smoking status, and duration of cough. Additionally, HIV status was determined either through self-reported diagnosis or confirmed positive test results. The dataset was split into training (n = 1,105) and validation (n = 1,038) subsets.

ICBHI: The ICBHI Respiratory Sound Database (Rocha et al., 2019) was originally compiled to support the International Conference on Biomedical Health Informatics (ICBHI) 2017 scientific challenge and is now publicly available for research. It consists of a combination of public and private datasets collected independently by two research teams across two different countries over several years. The dataset contains 5.5 hours of respiratory sound recordings, comprising 6,898 respiratory cycles from 126 subjects. The 920 audio samples in the dataset have been manually annotated by respiratory experts, classifying them based on the presence of crackles, wheezes, both, or no adventitious respiratory sounds. Additionally, the dataset provides diagnostic labels for chronic obstructive pulmonary disease (COPD), pneumonia, and asthma, enabling the development of machine-learning models for disease classification.

KAUH: The KAUH (King Abdulaziz University Hospital) dataset (Fraiwan et al., 2022) is a collection of respiratory sound recordings from 112 subjects, including 35 healthy individuals and 77 patients with pulmonary conditions. Lung sounds were recorded using a Electronic Stethoscope, which was placed at multiple points on the chest wall to capture respiratory sounds while avoiding heart sounds. The recordings were processed and extracted using Heart and Lung Sound Visualization software, which allows exporting data with three different filter settings (Bell, Diaphragm, and Extended) to emphasize different frequency ranges relevant to lung sounds.

B More Details of Audio Encoder

We utilized the Opera-CT encoder (Zhang et al., 2024a), to extract audio features from raw audio signals. Opera-CT is a contrastive learning-based hierarchical token-semantic audio transformer (Chen et al., 2022). It operates by dividing the mel-spectrogram into patches, which are embedded as input tokens for the transformer. The model leverages a hierarchical architecture with window attention, optimizing both computational efficiency and memory usage by restricting attention to localized windows. The transformer has 31 million parameters and produces output features of size $D_a = 768$.

C Baselines

We compared `RespiraMFM` with the following sate-of-the-art multimodal baselines:

Qwen2-Audio: Qwen2-Audio (Chu et al., 2024) is a large-scale multimodal language model capable of processing both speech and text inputs. It leverages the Qwen2 backbone with an integrated audio encoder that converts raw waveforms into latent representations aligned with textual embeddings through a shared multimodal transformer. For our experiments, we utilized the publicly released Qwen2-Audio model [2] and fine-tuned it on the respiratory sound classification task.

BTS: BTS (Kim et al., 2024) proposes a module called Bridging Text and Sound (BTS), which aligns respiratory audio and text metadata by utilizing CLAP (Elizalde et al., 2023) as a dual-purpose encoder for both modalities. In this approach, CLAP independently processes text and audio data through separate encoders. The resulting embeddings are then concatenated and passed through a linear classifier to perform the disease prediction.

RespLLM: RespLLM (Zhang et al., 2024b) introduces a multimodal approach using a pre-trained audio encoder and a Large Language Model for diagnosing respiratory diseases using audio recordings and patient metadata. RespLLM employs a trainable linear projector to align audio embeddings

[2] https://github.com/QwenLM/Qwen2-Audio

Table 5: Dataset-wise patient symptoms and medical history selection

Dataset	Patient information
Uk covid-19	Age, sex, cough, new continuous cough, runny or blocked nose, shortness of breath, sore throat, abdominal pain, diarrhea, fatigue, fever, headache, changes to sense of smell or taste, loss of taste, asthma, other symptoms
Coughvid	Age, sex, fever and muscle pain, other respiratory symptoms
TBscreen	Age, sex, fever, cough, night sweats, cough with blood, smoking status, previous TB history, HIV status, cough duration
ICBHI	Age, sex, BMI, child weight, child height, recording device placement
Coswara	Age, sex, cold, cough, diarrhea, fever, loss of smell and taste, muscle pain, breathing difficulties, fatigue, sore throat
CodaTB	Age, sex, fever, weight loss, night sweats, cough with blood, previous TB history, HIV status, cough duration
KAUH	Age, sex, recording device placement, sound type

with the language model's input space. In contrast, our method adopts a contrastively trained projection head, which enables more effective alignment between audio and text modalities.

D Additional Details on Contrastive Aligner

D.1 Model Architecture

The contrastive alignment module is implemented as a multi-layer perceptron (MLP) with normalization and regularization components. Specifically, the projection head maps an input embedding of dimension 768 into a higher-dimensional contrastive space of D_{LLM} through an intermediate hidden layer of size 1024. The architecture consists of a linear transformation followed by Layer Normalization, ReLU activation, and dropout (rate = 0.1). A final linear layer produces the output embeddings used for contrastive supervision.

D.2 Training

We trained the alignment module using the same dataset employed during instruction-tuning. The model was optimized for 100 epochs with learning rate 0.001.

E Additional Details on Instruction-Tuning

Table 6 presents the hyperparameter settings used in this work.

Hyperparameters	Value
Instruction tuning epochs	40
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.1
Total batch size	16
Maximum sequence length	256
Learning rate	2e-4
Learning rate optimizer	AdamW
Schedule	linear
Weight decay	0.1

Table 6: Training hyperparameters

F Additional Experiments

Ablation on Model Architecture: Table 7 presents the results across all tasks using various LLMs as a backbone. Among them, Phi-2 consistently outperforms other models, achieving the highest average score across all tasks.

Missing Input: To simulate a real-world scenario where some data may be missing, we evaluate `RespiraMFM` under various missing input conditions, as shown in Figure 9. This experiment is performed on task T5. Specifically, we remove one symptom from the patient metadata, restructure the prompt, and feed it to the model to calculate the AUC. The red dotted line represents the AUC when all symptoms are present, serving as a baseline. As observed in the plot, removing information such as gender does not significantly affect performance, with AUC either remaining the same or showing only a slight decrease. However, the removal of

Table 7: AUROC comparison for all respiratory disease recognition tasks using different LLMs as backbone. Here 'M' denotes the million level, and 'B' denotes the billion level. The heavy teal color indicates the highest results.

MODELS	SIZE	DIM	T1	T2	T3	T4	T5	T6	T7	T8	T9	Average
GPT2-Medium	345M	1024	0.911	0.688	0.773	0.826	0.918	0.508	0.713	0.605	0.677	0.735
LLaMA-3	1B	2048	0.906	0.645	0.767	0.996	0.913	0.563	0.686	0.547	0.619	0.738
Phi-2	2.7B	2560	0.906	0.648	0.817	0.998	0.904	0.670	0.768	0.617	0.800	0.792
LLaMA-3	8B	4096	0.905	0.676	0.796	0.851	0.853	0.711	0.642	0.594	0.617	0.738

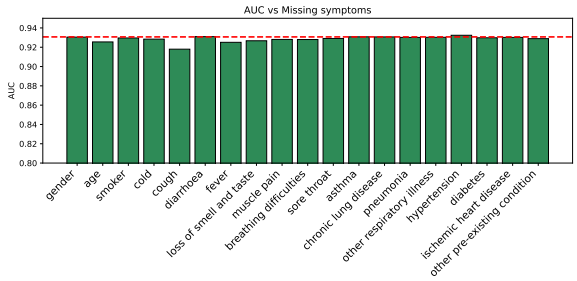


Figure 9: AUC on Task T5 under various missing input scenerios.The red dotted line represents the AUC when all symptoms are present.

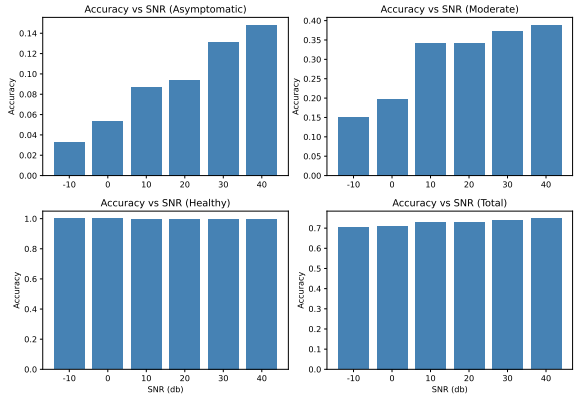


Figure 10: AUC on Task T5 under noisy condition.

symptoms like having cough results in a more noticeable drop in AUC compared to other symptoms, suggesting that these symptoms play a more critical role, which aligns with expectations in COVID detection. Although the AUC decreases when certain information is removed, the drop is not substantial, indicating that the system is sufficiently robust to handle missing inputs in real-world scenarios.

Experiment on Noisy Input:

To evaluate the robustness of `RespiraMFM` under noisy conditions, we conducted experiments by adding Gaussian noise to the audio inputs. We selected Task 5, corresponding to zero-shot evaluation on the Coswara dataset, as it provides both disease labels and detailed patient metadata, including health status and severity levels such as asymptomatic, moderate, and healthy. The results, shown in Figure 10, illustrate the model's performance across these conditions. We observe that under the asymptomatic setting where subtle acoustic cues are most critical, adding noise noticeably degrades performance. In contrast, for the healthy group, the addition of noise has minimal impact, indicating that the model remains relatively robust in scenarios with less disease-related acoustic variability.