Zero-Shot Respiratory Sound Classification through LLM-Augmented Audio-Text Alignment

Mustafa Talha İlerisoy[1], *Hung Manh Pham*[2], *Mathias Funk*[1], *Mykola Pechenizkiy*[1], *Aaqib Saeed*[1]

[1] Eindhoven University of Technology, Eindhoven, Netherlands
[2] Singapore Management University, Singapore

{m.t.ilerisoy, m.funk, m.pechenizkiy, a.saeed}@tue.nl, hm.pham.2023@phdcs.smu.edu.sg,
Github: https://github.com/mtilerisoy/REACH

Abstract

Self-supervised respiratory encoders lack semantic grounding in clinical domain needed for zero-shot inference, limiting their utility without task-specific labeled data. We propose a framework that aligns these encoders with medical terminology in a shared latent space turning them into a zero-shot-capable foundation model. To address paired data scarcity, we use a medical LLM to synthesize structured reports from metadata, creating dense semantic anchors for contrastive learning. Our training combines a sigmoid-based contrastive loss with encoder's native SSL objective and similarity-aware negative sampling to sharpen pathological boundaries. Across 9 tasks on 6 datasets, our method achieves a 61.3% mean zero-shot AUC, surpassing CLAP (51.4%) and Qwen2-Audio (54.9%) while reaching the highest linear probing AUC (71.6%) with only 43% of data used by full-scale baselines, showing that structured semantic alignment outperforms large-scale, general-purpose models in clinical diagnostics.

Index Terms: Respiratory Foundation Model, Multimodal Alignment, Zero-shot Learning, Clinical Diagnostics

1. Introduction

Respiratory auscultation remains one of the most widely practiced diagnostic procedures worldwide [1], yet its interpretation is highly subjective and dependent on clinical expertise that is unevenly distributed [2]. AI-driven auscultation tools offer a path toward scalable, objective screening [3, 4, 5], particularly in low-resource settings where pulmonologists are unavailable [6, 7]. For such tools to be clinically viable, however, they must generalize to novel pathologies *without* task-specific labeled data, a capability known as zero-shot inference, which remains largely unattained in respiratory audio modeling [8].

Self-supervised learning has produced powerful acoustic encoders that form the backbone of current respiratory AI. General-purpose models [9, 10] learn spectro-temporal representations through masked autoencoding on broad audio corpora, while domain-specific efforts [11] tailor these to respiratory events via large-scale pre-training. These encoders function as *acoustic experts*, resolving fine-grained auscultatory patterns (crackles, wheezes, stridor) with high fidelity. Yet, their embedding spaces remain *semantically opaque*: clusters corresponding to distinct pathologies may be linearly separable, but neither is anchored to the medical concept it encodes. Consequently, every new clinical task requires supervised fine-tuning with scarce, expert-curated annotations [12], creating a bottleneck that limits real-world deployment.

Multimodal contrastive learning addresses this by aligning audio and text in a shared space, enabling zero-shot classification via natural language prompts that are validated for vision [13] and general audio [14]. However, on clinical respiratory benchmarks, general-purpose audio-text models fail to outperform unimodal baselines [11, 15]: they are grounded in everyday captions and lack clinical language. Training a medical audio-text model from scratch would require paired audio-report datasets that do not exist at scale for respiratory sounds [16].

We observe that strong acoustic features and clinical language understanding already exist in isolation: respiratory encoders [11] capture the diagnostic signal, while medical text encoders [17] capture disease-level semantics. The core problem is not representation learning but *representation alignment* [18]. While other methods align vision-language spaces [19] or employ generative instruction-tuning for respiratory health [20], we address paired-data scarcity through sample-efficient contrastive alignment. Building on this, we propose REACH (REport-Augmented Contrastive alignment for respiratory Health), a **semantic alignment framework** that re purposes pre-trained unimodal respiratory encoders into zero-shot capable multimodal tools through targeted post-training. Our approach treats existing encoders as complementary assets and focuses on learning the bridge between them. The design is *modular*, accepting any transformer-based audio backbone; *data efficient*, requiring no paired audio report corpora; and *non-destructive*, preserving acoustic fidelity throughout.

Our framework addresses three challenges: (1) absent paired data; we use an off-the-shelf medical-grade LLM [21] to synthesize reports from metadata, creating semantic anchors for contrastive learning; (2) low textual diversity; we employ FAISS-based [22] similarity-aware negative sampling to mine distant negatives; and (3) catastrophic feature degradation; we combine a sigmoid contrastive loss [23] with the encoder's native reconstruction objective as a structural regularizer. The text encoder [17] remains frozen as a fixed semantic reference. Across 9 tasks on 6 datasets, our method achieves 61.3% mean zero-shot AUC, surpassing CLAP (51.4%) and Qwen2-Audio (54.9%), while attaining the highest linear probing AUC (71.6%) with only 43% of the training data used by the full-scale baseline [11].

The text branch employs a medical text encoder [17] that remains fully frozen during alignment as a fixed semantic reference; only the audio branch and projection heads are optimized. We evaluate on 9 tasks from 6 publicly available respiratory datasets spanning in-domain and out-of-domain settings. Our method achieves a mean zero-shot AUC of 61.3%, surpassing CLAP (51.4%) and the 7B parameter Qwen2 Audio (54.9%), while using only 43% of the pre-training data available to the full-scale baseline [11]. Importantly, alignment does not compromise unimodal capability: our model attains the highest mean linear probing AUC (71.6%), exceeding even the baseline

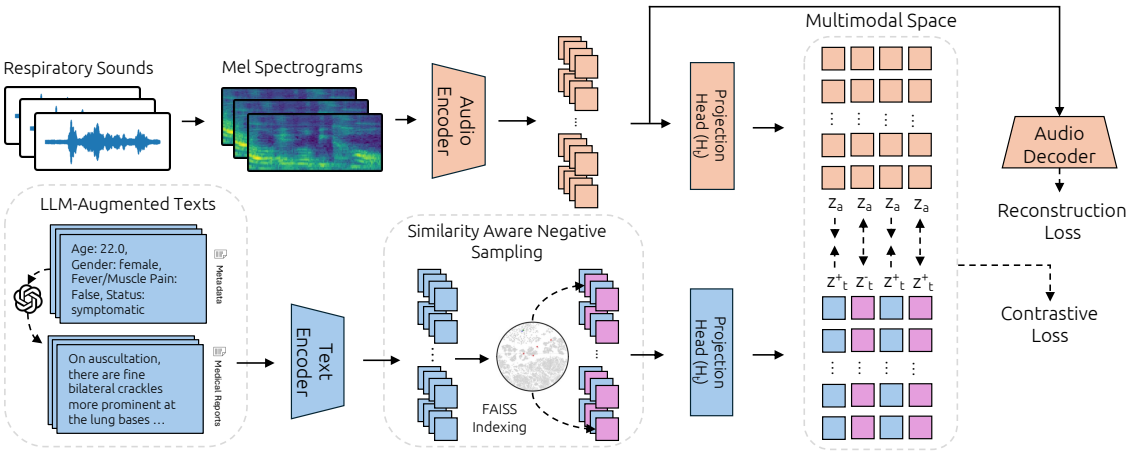

Figure 1: *Overview of REACH: a semantic alignment framework that repurposes pre-trained unimodal encoders for zero-shot respiratory sound classification via LLM-augmented report synthesis, similarity-aware negative sampling, and structure-preserving contrastive alignment.*

trained on the complete proprietary corpus. Our contributions are as follows:

1. A **modular semantic alignment framework** (REACH) that transforms pre-trained unimodal respiratory encoders into zero-shot capable multimodal models, decoupling acoustic pre-training from clinical language grounding.
2. An **LLM augmented report synthesis** pipeline that converts discrete metadata into clinically structured text, creating semantic anchors that eliminate the need for paired audio report datasets.
3. A **structure-preserving alignment strategy** that reconciles cross-modal transfer with preservation of acoustic features by jointly optimizing a sigmoid contrastive objective [23] with the encoder's native reconstruction loss, complemented by similarity-aware negative sampling.
4. **Comprehensive empirical validation** across 9 tasks on 6 datasets, demonstrating that targeted alignment with 57% less data outperforms both full-scale unimodal pre-training and general-purpose audio language models of significantly larger capacity.

2. Methodology

2.1. Problem Formulation

Let f_a denote a pre-trained unimodal audio encoder and f_t a pre-trained text encoder, producing embeddings $\mathbf{x}_a \in \mathbb{R}^m$ and $\mathbf{x}_t \in \mathbb{R}^n$, respectively. These embedding spaces are independently learned and share no semantic correspondence. Our objective is to learn projection heads $H_a : \mathbb{R}^m \to \mathbb{R}^d$ and $H_t : \mathbb{R}^n \to \mathbb{R}^d$ that map both modalities into a shared space $\mathbb{R}^d$ where semantically corresponding pairs are aligned.

Let $\mathbf{z}_a = H_a(\mathbf{x}_a)$ and $\mathbf{z}_t^+ = H_t(\mathbf{x}_t^+)$ denote projections of a matched pair and $\mathbf{z}_t^- = H_t(\mathbf{x}_t^-)$ a non-corresponding report. We seek projections satisfying:

$$S(\mathbf{z}_a, \mathbf{z}_t^+) > S(\mathbf{z}_a, \mathbf{z}_t^-) \quad \forall\, \mathbf{x}_t^- \neq \mathbf{x}_t^+ \tag{1}$$

where $S(\cdot, \cdot)$ denotes cosine similarity. For instance, the projection of a recording with prominent wheezes should be closer to a report describing *"expiratory wheezes in a 70-year-old female"* than to one describing *"clear breath sounds with no adventitious findings."* Once aligned, zero shot classification reduces to $\arg\max_c\ S(\mathbf{z}_a, \mathbf{z}_{t_c})$ over class specific text anchors $\{\mathbf{z}_{t_c}\}_{c=1}^{C}$, each generated by encoding a clinical prompt for class c.

2.2. Semantic Anchor Generation via LLM Augmented Report Synthesis

Contrastive alignment requires semantically rich, paired text for each audio sample. Existing respiratory datasets provide only discrete metadata: sound type (e.g., cough, breathing cycle), adventitious sound labels (e.g., wheeze, crackle), recording location (e.g., posterior lower lobe), patient demographics, and disease diagnosis. These categorical fields lack the narrative structure needed to serve as effective linguistic anchors.

We employ a medical-grade off-the-shelf LLM (i.e., GPT-4 [21]) to transform this metadata into standardized clinical reports, conditioned on a prompt instructing the model to adopt the role of a pulmonologist:

> *You are a Pulmonologist tasked with interpreting respiratory auscultation findings. Based on the given conditions, write 2–3 lines covering clinically relevant information. Only use the information given to write about conditions. Do NOT mention anything about further evaluation or characterization.*

This prompt enforces two constraints: restricting the LLM to the provided metadata prevents hallucination of ungrounded clinical details, and prohibiting follow-up recommendations keeps the anchors tightly coupled to the observable acoustic content. To prevent information leakage, the metadata used during alignment is strictly partitioned from evaluation data; at inference, the model encounters clinical concepts (e.g., *COPD*) never paired with audio during training.

2.3. Model Architecture

Our framework is *architecturally modular*, accepting any pair of transformer-based unimodal encoders. We instantiate it with two backbones selected for complementary strengths.

Audio Encoder. We adopt a respiratory-specific transformer [11] pre-trained via masked spectrogram reconstruction on a large-scale mel spectrogram corpus (see Table 2 for the specific variant), representing the current state of the art for respiratory feature extraction.

Text Encoder. We employ the text branch of MedSigLIP [17], a medical vision language encoder pre-trained on paired chest radiographs and clinical reports via sigmoid contrastive loss. Its representations encode clinical nomenclature and report structure, making it an ideal fixed semantic reference. The vision encoder is discarded; only the text encoder is retained. Notably, this encoder has never seen respiratory audio, so all cross-modal alignment is learned through our framework.

Projection Heads. We introduce lightweight projection heads H_a and H_t, each consisting of a linear layer followed by layer normalization, mapping m and n dimensional features into the shared d dimensional space.

2.4. Alignment Objective

We optimize a dual objective that drives cross-modal alignment while preserving the encoder's pre-trained feature geometry:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{MSE}} \tag{2}$$

Contrastive Alignment ($\mathcal{L}_{\text{contrastive}}$). We adopt a SigLIP-based [23] formulation that treats each audio text pair as an independent binary classification problem. For a batch of N pairs:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\sigma\left(y_{ij}\left(\alpha \cdot S(\mathbf{z}_a^i, \mathbf{z}_t^j)\right)\right) \tag{3}$$

where $y_{ij} = +1$ if (i, j) is a matched pair and $y_{ij} = -1$ otherwise, σ is the sigmoid function, and α is a learnable temperature parameter. Unlike softmax-based InfoNCE [13], this formulation requires no global batch normalization, making it robust to the small batch sizes typical in medical settings.

Masked Reconstruction Regularizer ($\mathcal{L}_{\text{MSE}}$). Contrastive fine tuning alone risks distorting the pre-trained acoustic manifold [24]. We retain the encoder's original self-supervised objective: during each forward pass, random spectrogram patches are masked, and the encoder reconstructs them. This MSE loss acts as a *structural regularizer*, anchoring representations to their pre-trained geometry while the contrastive term reshapes the space for cross-modal compatibility.

2.5. Similarity Aware Negative Sampling

In clinical domains, semantically distinct conditions can produce textually similar reports (e.g., "wheezes in a 65-year-old male with asthma" vs. "wheezes in a 60-year-old male with COPD"), rendering in batch random negatives insufficiently contrastive.

Offline Indexing. Before training, we encode the entire report corpus with the frozen text encoder and construct a FAISS index [22] over the resulting embeddings.

Online Negative Swapping. During each training step, 50% of audio samples in the batch undergo negative swapping: we query the FAISS index to retrieve the k^{th} furthest embedding (k=10) from the positive anchor, which replaces the in-batch negative. Selecting k=10 rather than the absolute furthest point avoids degenerate outlier negatives while ensuring substantial semantic distance. This forces the model to maximize the margin between unrelated clinical pathologies in $\mathbb{R}^d$.

2.6. Training Strategy

The text encoder remains fully frozen; only the audio encoder and projection heads are updated (lr = 1×10^{-5}, 100 epochs). This asymmetry ensures the acoustic manifold migrates toward the stable clinical text space, and that inference prompts are interpreted in the same semantic frame used during training.

3. Experimental Setup

3.1. Evaluation Benchmark

We evaluate on a comprehensive benchmark [11] comprising 6 publicly available respiratory sound datasets to evaluate representation quality, generalization, and cross modal alignment.

Table 1: *Summary of curated datasets; shaded rows denote OOD datasets withheld from pre-training and alignment.*

Dataset	ID	Task	Modality	#Samples
CovidUK [25]	T1	Covid/Non-covid	Exhalation	840 / 1660
	T2	Covid/Non-covid	Cough	840 / 1660
COUGHVID [26]	T3	Covid/Non-covid	Cough	547 / 5628
	T4	Female/Male	Cough	2468 / 4795
ICBHI [27]	T5	COPD/Healthy	Lung sounds	793 / 35
Coswara [28]	T6	Smoker/Non-smoker	Cough	201 / 747
	T7	Female/Male	Cough	759 / 1737
KAUH [29]	T8	Obstructive/Healthy	Lung sounds	129 / 105
Resp.@TR [30]	T9	5-class COPD severity	Lung sounds	72/60/84/84/204

The benchmark consists of 9 tasks grouped into 5 *in domain* (ID) tasks, drawn from datasets used during pre training and alignment, and 4 *out of domain* (OOD) tasks, drawn from datasets the model has never encountered in any training phase. All tasks are binary classification problems except T9, which is a five class COPD severity grading task. Table 1 summarizes the dataset characteristics, task definitions, and class distributions.

3.2. Evaluation Protocol

We evaluate along three axes. Linear probing: a frozen encoder plus single linear layer (5-seed average). kNN: non-parametric classification measuring embedding geometry directly. Zero-shot: each class is represented by a clinically descriptive text prompt encoded into the shared space; classification assigns each sample to the nearest text anchor by cosine similarity. We report AUC (%) throughout.

3.3. Baselines

We organize baselines into three categories to contextualize our results against models with fundamentally different capabilities: **Unimodal encoders** (linear probing and kNN only): OpenSMILE [31], VGGish [10], AudioMAE [9], and the OPERA family [11] (OCT, OCE, OGT). **General-purpose audio-text**: CLAP [14] (all three protocols). **Audio-language decoders** (zero-shot only): Qwen2-Audio 7B [32] and Audio-Flamingo-3 [33], included to test whether scale compensates for domain alignment.

3.4. Data Fairness and Training Conditions

A critical consideration in our evaluation is data parity. The full OPERA training corpus contains datasets with varying access conditions: a subset is publicly available, while the remainder requires institutional data use agreements. To ensure a reproducible and fair comparison, we retrain OPERA variants (denoted † in Table 2) using only the openly accessible subset, comprising approximately 43% of the original training volume. Our alignment framework operates on this same subset, ensuring all comparisons against † variants are data-matched and isolate the effect of our alignment strategy from differences in training scale. We additionally report full-corpus OPERA results (without †) to contextualize absolute performance, but exclude these from the primary ranking due to the differing training data.

4. Results

Table 2 presents the full benchmark results. We analyze them along three axes: preservation of unimodal capability, improvement in latent space structure, and zero-shot cross-modal transfer.

Alignment enhances unimodal features, rather than degrading them. Restricting the acoustic encoder [11] to the openly accessible 43% of its original training corpus incurs a 2.3 point drop in mean linear probing AUC (67.7 → 65.4). Our alignment

Table 2: *Benchmark results (% AUC).* † *: trained on open access data only (≈43%).* *Shaded* *: full corpus, excluded from ranking.* Δ*: gain over OGT*†*.* ***Bold*** *and* *underline* *indicates the best and the second best respectively.*

	In Domain					OOD				
	T1	T2	T3	T4	T5	T6	T7	T8	T9	Avg
Linear Probing ↑										
CLAP	56.5	64.8	**59.9**	66.5	**93.3**	**68.0**	74.2	**69.7**	63.6	68.5
OpenSMILE	55.0	64.9	53.7	67.7	57.9	53.4	75.3	63.6	49.4	60.1
VGGish	58.0	55.7	53.8	60.0	60.5	50.7	60.6	60.5	59.0	57.6
AudioMAE	54.9	61.6	55.4	62.8	88.6	54.9	72.4	61.6	51.0	62.6
OCT	58.6	70.1	57.8	79.5	85.5	68.5	87.4	72.2	62.5	71.3
OCE	55.1	62.9	56.6	72.1	87.2	67.4	80.1	74.1	68.3	69.3
OGT	60.5	67.7	55.2	73.5	74.1	65.0	82.5	70.3	60.6	67.7
OCT†	58.7	67.2	56.7	76.9	88.0	67.4	84.4	68.4	66.2	70.4
OCE†	55.1	65.4	54.8	70.6	87.1	64.2	75.0	63.0	65.2	66.7
OGT†	56.2	64.7	55.6	70.3	74.5	62.9	77.8	64.7	61.7	65.4
REACH (Ours)	**63.8**	**69.8**	56.3	**78.4**	90.2	66.1	**86.3**	66.9	**66.6**	**71.6**
Δ	7.6	5.1	0.7	8.1	15.7	3.2	8.5	2.2	4.9	6.2
k-NN ↑										
CLAP	54.4	57.9	51.8	59.0	89.0	53.3	61.4	**71.5**	57.3	61.7
OpenSMILE	49.7	51.3	47.1	52.2	77.7	52.2	51.3	71.3	55.2	56.4
VGGish	53.2	53.9	50.3	56.8	81.6	51.3	61.3	53.9	47.3	56.6
AudioMAE	52.4	56.5	48.4	52.2	89.5	51.6	58.0	58.7	53.1	57.8
OCT	48.7	60.9	55.2	71.0	62.7	65.8	75.9	66.6	55.4	62.5
OCE	49.3	53.8	51.7	60.6	67.4	56.2	72.1	65.8	57.4	59.3
OGT	51.3	55.6	50.9	60.7	66.0	57.1	68.5	59.4	49.0	57.6
OCT†	48.6	59.3	**56.7**	**68.8**	74.4	61.5	72.9	57.0	52.4	61.3
OCE†	48.9	56.2	50.0	62.1	73.3	57.5	67.6	57.1	54.7	58.6
OGT†	50.1	56.5	52.3	61.4	68.3	56.1	72.4	58.3	53.8	58.8
REACH (Ours)	**55.5**	**62.4**	52.4	67.2	**90.1**	**64.4**	**77.0**	56.9	**60.6**	**65.2**
Δ	5.6	5.9	0.1	5.8	21.8	8.3	6.4	−1.4	6.8	6.4
Zero Shot ↑										
CLAP	52.4	**62.6**	43.6	63.4	51.8	48.6	61.5	31.3	47.5	51.4
Qwen2 Audio 7B	52.0	55.0	47.8	63.1	75.0	46.7	56.8	46.9	51.2	54.9
Flamingo 3	48.7	53.7	51.1	68.9	25.8	53.4	**71.8**	42.9	**65.1**	53.5
REACH (Ours)	**58.7**	61.5	**51.3**	**67.1**	**76.7**	**67.4**	65.0	**52.0**	52.1	**61.3**

framework not only recovers this loss but also surpasses the full corpus baseline, reaching a 71.6 mean AUC (+6.2 over OGT†, +3.9 over the original OGT). Gains span all tasks in comparison with OGT†, with the largest improvements on T5 (+15.7), T7 (+8.5), T4 (+8.1), and T1 (+7.6). This indicates that the semantic anchors introduced during alignment provide a complementary training signal that enriches acoustic representations beyond what additional audio data alone achieves.

Contrastive alignment reorganizes the audio manifold into clinical clusters. The kNN protocol measures embedding geometry without learned classification parameters. Our method achieves a 65.2 mean AUC, outperforming all baselines, including OGT† (58.8) by 6.4 points. The strongest gain appears on T5 (COPD/Healthy: 90.1 vs. 68.3), where medical text anchors provide the semantic scaffolding to separate pathological from healthy lung sounds. The only regression is T8 (-1.4), which we attribute to its small sample size (234 samples), insufficient to form stable neighborhoods in the projected space.

Targeted medical alignment outperforms both general-purpose and large-scale models at zero-shot inference. The zero-shot setting is the centerpiece of our evaluation: the model must classify audio solely via cosine similarity to text anchors, with no labeled data. OPERA baselines [11] cannot operate here due to their unimodal nature. Our method achieves 61.3% mean AUC, substantially exceeding CLAP (51.4%) and Qwen2 Audio 7B (54.9%). CLAP's failure despite being designed for zero-shot audio classification confirms that general-purpose text grounding is insufficient for clinical domains. That a 7B parameter generative model also falls short demonstrates that scale alone cannot substitute for domain-specific semantic alignment. Among individual tasks, strong results on T6 (67.4%) and T5 (76.7%) reflect pathologies with distinct auscultatory signatures that map well to clinical descriptions. Performance on T3 (51.3%) and T9 (52.1%) remains near chance, reflecting the difficulty of distinguishing subtle cough variations (T3) and grading severity across five classes from text alone (T9).

Table 3: *Ablation results (mean % AUC across 9 tasks).*

Configuration	Lin. Probe ↑	k-NN ↑	Zero Shot ↑
Audio Encoder from Scratch	66.0	58.8	55.1
Random Negatives	70.6	63.9	53.5
No $\mathcal{L}_{\text{MSE}}$	71.1	62.8	54.1
Frozen Audio Backbone	66.9	58.5	48.8
BERT Text Encoder	65.1	59.6	49.9
Negative Sampling K=100	69.9	61.1	54.2
REACH (Ours)	**71.6**	**65.2**	**61.3**

4.1. Ablation Studies

Table 3 isolates each component's contribution. A consistent pattern emerges: several ablations maintain linear probing while severely degrading zero-shot AUC, indicating the removed component is essential for cross-modal alignment, not unimodal feature quality.

Training the audio encoder from scratch yields near-random zero-shot AUC (55.1) despite adequate linear probing (66.0), confirming that alignment requires an established acoustic manifold to reorganize. Replacing FAISS-based distant negatives with random in-batch sampling preserves linear probing (70.6) but collapses zero-shot AUC to 53.5, as clinical reports for related pathologies are near-paraphrases that only distant negatives can disambiguate. Removing $\mathcal{L}_{\text{MSE}}$ produces the most revealing dissociation: a 7.2-point zero-shot drop (54.1) with a only 0.5-point linear probing decrease, showing that without reconstruction regularization, the contrastive objective distorts pre-trained geometry into a space no longer coherent for text anchor mapping. Freezing the audio backbone and training only projection heads yields the weakest zero-shot AUC (48.8), confirming that partial encoder adaptation is necessary. Substituting MedSigLIP with BERT drops zero-shot AUC by 11.4 points (49.9%), confirming general-purpose text encoders lack clinical grounding for effective alignment. Increasing negative sampling to $K{=}100$ also degrades zero-shot AUC (54.2%), indicating similar negatives fail to provide clear boundaries between unrelated pathologies.

5. Conclusion

We introduced REACH, a framework for aligning pre-trained respiratory audio encoders with clinical text representations, enabling zero-shot classification without paired audio-report data. Our results suggest that the primary barrier to zero-shot respiratory diagnostics is not acoustic representation quality, which existing self-supervised encoders already capture well, but rather their disconnection from clinical semantics. By treating an off-the-shelf LLM as a metadata-to-report translator, even discrete categorical annotations can be elevated into rich linguistic anchors sufficient for contrastive alignment. Most notably, our framework surpasses baselines trained on 57% more data and models with orders-of-magnitude more parameters, suggesting that targeted semantic bridging of mature unimodal models is a more sample-efficient paradigm than scaling pre-training or building multimodal architectures from scratch. We believe this principle generalizes beyond respiratory audio: any clinical domain with capable unimodal encoders and structured metadata is a candidate for the same alignment strategy.

6. Generative AI Use Disclosure

Generative AI tools were used solely for language editing and polishing to improve clarity and readability of the manuscript. All technical content, experimental design, analysis, and conclusions were created by the authors. The authors take full responsibility for the content of this paper.

7. Acknowledgments

This work was supported by the NWO AiNed Fellowship Grant of A.S., and in part by Google.org and the Google Cloud Research Credits program through the Gemini Academic Program. We also acknowledge the use of the Dutch National Supercomputer Snellius for essential computational tasks.

8. References

[1] J. Cook, M. Umar, F. Khalili, and A. Taebi, "Body acoustics for the non-invasive diagnosis of medical conditions," *Bioengineering*, vol. 9, no. 4, p. 179, Apr. 2022.

[2] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clin. Med. Circ. Respir. Pulm. Med.*, vol. 2, pp. CCRPM–S530, 2008.

[3] E. P. Doheny, B. P. O'Callaghan, V. S. Fahed, J. Liegey, C. Goulding, S. Ryan, and M. M. Lowery, "Estimation of respiratory rate and exhale duration using audio signals recorded by smartphone microphones," *Biomedical Signal Processing and Control*, vol. 80, p. 104318, 2023.

[4] A. Moberg, Ö. Ingi Emilsson, S. Hansdottir, T. Asmundsson, A. Malinovschi, H. Melbye, and D. Ludviksdottir, "Lung auscultation - today and tomorrow- a narrative review," *Expert Rev. Respir. Med.*, vol. 19, no. 8, pp. 879–885, Aug. 2025, epub 2025 May 26. PMID: 40415608.

[5] D. Drummond, I. Adejumo, K. Hansen, V. Poberezhets, G. Slabaugh, and C. Y. Hui, "Artificial intelligence in respiratory care: perspectives on critical opportunities and challenges," *Breathe (Sheff.)*, vol. 20, no. 3, p. 230189, Dec. 2024, pMID: 39660082; PMCID: PMC11629173.

[6] F. Liu, Z. Li, Q. Yin *et al.*, "A multimodal multidomain multilingual medical foundation model for zero shot clinical diagnosis," *npj Digit. Med.*, vol. 8, no. 1, p. 86, Jan. 2025.

[7] N. Hasani, F. Farhadi, M. A. Morris, M. Nikpanah, A. Rhamim, Y. Xu, A. Pariser, M. T. Collins, R. M. Summers, E. Jones, E. Siegel, and B. Saboury, "Artificial intelligence in medical imaging and its impact on the rare disease community: Threats, challenges and opportunities," *PET Clin.*, vol. 17, no. 1, pp. 13–29, Jan. 2022, pMID: 34809862; PMCID: PMC8764708.

[8] M. Rath, J. Coetzee, M. van Breda, and B. van Breda, "The development and evaluation of ai-based tuberculosis screening with a digital stethoscope used to capture lung sounds. a case-control study," *medRxiv*, 2025. [Online]. Available: https://www.medrxiv.org/content/early/2025/07/31/2025.07.31.25332442

[9] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, Dec. 2022, pp. 28 708–28 720.

[10] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[11] Y. Zhang, T. Xia, J. Han, Y. Wu, G. Rizos, Y. Liu, M. Mosuily, J. Chauhan, and C. Mascolo, "Towards open respiratory acoustic foundation models: Pretraining and benchmarking," *arXiv preprint arXiv:2406.16148*, Nov. 2024.

[12] A. Berger, T. A. Lagones, L. Grigull, L. Fendrich, T. Bell, H. Högl, G. Ernst, R. Schmidt, D. Bascom, R. Sifa, and M. Lübbering, "Tackling data sparsity and combinatorial challenges in rare disease matching with medical informed machine learning," *2024 IEEE International Conference on Big Data (BigData)*, pp. 6430–6438, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:275586197

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[14] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[15] D. Niizumi, D. Takeuchi, M. Yasuda, B. T. Nguyen, Y. Ohishi, and N. Harada, "Towards pre-training an effective respiratory audio foundation model," in *Proc. Interspeech 2025*, Aug. 2025, pp. 998–1002.

[16] S. Shokouhmand, S. Bhatt, and M. Faezipour, "Artificial intelligence in respiratory health: A review of AI-driven analysis of oral and nasal breathing sounds for pulmonary assessment," *Electronics*, vol. 14, no. 10, p. 1994, 2025.

[17] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.

[18] R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang, "Understanding multimodal contrastive learning and incorporating unpaired data," in *Proc. 26th Int. Conf. Artif. Intell. Statist. (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 206. PMLR, Apr. 2023, pp. 4348–4380. [Online]. Available: https://proceedings.mlr.press/v206/nakada23a.html

[19] J. Yang, Z. Wu, Y. Zhao, and Y. Ma, "Language-image alignment with fixed text encoders," 2025. [Online]. Available: https://arxiv.org/abs/2506.04209

[20] Y. Zhang, T. Xia, A. Saeed, and C. Mascolo, "Respllm: Unifying audio and text with multimodal llms for generalized respiratory health prediction," in *Proceedings of the 4th Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research, vol. 259. PMLR, 15–16 Dec 2025, pp. 1053–1066. [Online]. Available: https://proceedings.mlr.press/v259/zhang25a.html

[21] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[22] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024.

[23] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 941–11 952.

[24] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022.

[25] H. Coppock, G. Nicholson, I. Kiskin, V. Koutra, K. Baker, J. Budd, R. Payne, E. Karoune, D. Hurley, A. Titcomb *et al.*, "Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers," *Nat. Mach. Intell.*, pp. 229–242, 2024.

[26] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 156, 2021.

[27] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome,

A. Marques, N. Maglaveras, R. P. Paiva, I. Chouvarda, and P. de Carvalho, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological Measurement*, vol. 40, p. 035001, Mar 2019.

[28] D. Bhattacharya *et al.*, "Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection," *Scientific Data*, vol. 10, no. 397, 2023.

[29] M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian, "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," *Data in Brief*, vol. 35, p. 106913, 2021.

[30] G. Altan, Y. Kutlu, Y. Garbi, A. O. Pekmezci, and S. Nural, "Multimedia respiratory database (respiratorydatabase@tr): Auscultation sounds and chest x-rays," *Natural and Engineering Sciences*, vol. 2, no. 3, p. 59–72, 2017.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[32] J. B. et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[33] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S. gil Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," 2025. [Online]. Available: https://arxiv.org/abs/2507.08128