# LungTTA: Text-to-Audio Generation of Synthetic Lung Sounds for Respiratory Health

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Respiratory audio analysis is still limited by data scarcity, as real recordings are difficult to collect and often involve privacy and clinical constraints, which makes it harder to train robust machine learning models. We introduce LungTTA, a text-to-audio framework based on a latent diffusion model, which generates respiratory sounds such as cough, breathing, and phonation from structured prompts. The model is fine-tuned on 116,660 publicly available recordings and includes a retrieval-based memory component together with watermarking for traceability. We evaluate the generated audio using Fréchet Audio Distance (FAD), Kullback–Leibler (KL) divergence, and Inception Score (IS), and also introduce PRISM (Pulmonary Respiratory Integrity & Similarity Metric) a domain aware metric designed to capture respiratory signal structure. LungTTA achieves a FAD of 2.72, KL of 0.50, IS of 1.22, and PRISM of 0.23, compared to Stable Audio Open (6.73, 0.67) for FAD and KL, Make-An-Audio (1.54) for IS, and RespAgent (0.24) for PRISM. In human evaluation, LungTTA achieves 80.91 (Overall Quality, OVL) and 75.13 (Relevance to Text, REL), compared to RespAgent (59.27, 58.97) and EZAudio (55.24, 52.69), while expert assessment yields 58.33 (OVL), 44.44 (REL), and 38.89 (Clinical Relevance for Assessment, CRA), compared to RespAgent (56.94, 43.06, 36.11) and EZAudio (36.11, 29.17, 33.33). In a downstream COVID-19 cough classification task, LungTTA improves performance under a VGGish-based setting, increasing AUC from 0.7331 (no augmentation) and 0.7631 (classical augmentation) to 0.7701 using LungTTA. These results demonstrate that LungTTA-generated synthetic respiratory audio can be used as an effective data augmentation method.

**Keywords:** text-to-audio generation, respiratory sound synthesis, latent diffusion models, data augmentation, pulmonary health, synthetic medical audio

## 1 Introduction

Pulmonary diseases such as Chronic Obstructive Pulmonary Disease (COPD) and asthma remain among the leading causes of morbidity and mortality worldwide, motivating the development of reliable tools for diagnosis and continuous monitoring (Soriano & et al, 2017; Li et al., 2020). Audio-based approaches, including lung-sound analysis and machine learning classification, have shown that coughs, breathing sounds, and phonation signals can be used as acoustic biomarkers for respiratory health assessment (Mosuily et al., 2023; Nemati et al., 2022). However, collecting high-quality respiratory audio data is constrained by privacy and ethical requirements, dependence on clinical recording equipment, and the difficulty of recruiting representative patient cohorts. As a result, many publicly available datasets are limited in both size and diversity, which restricts the robustness and generalization of downstream machine learning models (Xia et al., 2022). Synthetic audio generation has therefore been explored as a way to mitigate data scarcity. Recent text-to-audio (TTA) models such as AudioLDM2 and Stable Audio Open generate high-quality general-purpose audio conditioned on text descriptions (Liu et al., 2023; Majumder et al., 2024; Evans et al., 2025). However, these models are trained in generalized audio domains and do not explicitly model the temporal structure, frequency characteristics, or task-specific requirements of respiratory sounds. Conventional augmentation methods, such as time shifting, noise injection, and filtering, operate directly on existing recordings and

therefore tend to produce variations of the same signals rather than genuinely new respiratory patterns (Xia et al., 2022). GAN-based approaches attempt to address this by generating new samples, but in practice they can be difficult to train and may produce unstable or inconsistent outputs in respiratory audio settings (Chakraborty et al., 2024). More recently, diffusion-based models have been adopted as a more stable alternative, with the ability to model complex audio distributions at higher fidelity (Feng et al., 2024). We introduce LungTTA, a text-to-audio generative framework for synthetic respiratory sound generation. LungTTA uses prompt-based conditioning to control attributes such as sound type, age, and smoking status, and is trained on a curated collection of publicly available respiratory datasets spanning cough, breathing, vowel phonation, and speech-based tasks. The goal is not to replace clinical data collection, but to provide a controlled and reproducible method for generating additional training data in data-scarce settings. All generated audio is embedded with a digital watermark that enables identification of synthetic samples.

We summarize the specific contributions of this work below.

- We present **LungTTA**, a domain-specific text-to-audio framework for respiratory sound synthesis, with a unified conditioning formulation $Z_{\text{cond}} = [H; Z_{\text{mem}}; Z_{\text{meta}}]$ that integrates prompt semantics, retrieval-based exemplar priors, and structured metadata.
- We formulate a metadata-grounded prompt construction and retrieval-guided conditioning pipeline that enables reproducible control of clinically relevant attributes.
- We introduce **PRISM (Pulmonary Respiratory Integrity & Similarity Metric)**, a domain-aware evaluation metric that complements standard metrics (FAD, KL, IS) by quantifying respiratory signal structure using waveform-level features.
- We perform evaluation using objective metrics, human and expert listening studies, and a downstream COVID-19 cough classification task, showing that LungTTA improves synthetic-data quality and downstream augmentation performance under controlled experimental settings.

## 2 Related Work

Respiratory audio modeling remains challenging due to limited datasets and variability in recording conditions, devices, and annotation quality. Although such variability can in principle support generalization, inconsistencies across datasets often make it difficult to develop robust and reliable models in practice (Xia et al., 2022; Niizumi et al., 2025). In addition, reliance on real patient recordings introduces privacy, ethical, and logistical constraints, which limit large-scale data collection and reinforce data scarcity as a key bottleneck in this domain. Early work addressed these limitations through classical augmentation techniques, applying signal-level transformations such as time shifting, noise injection, and filtering to expand training data. These methods are useful for regularization, but they primarily generate variations of existing recordings and provide limited control over clinically meaningful respiratory structure (Xia et al., 2022). GAN-based methods, such as CoughGAN Ramesh et al. (2020), introduced learned synthesis of respiratory sounds, but in practice they are often affected by unstable training, mode collapse, and reduced diversity, which limits their suitability for high-fidelity medical audio generation (Chakraborty et al., 2024). More recently, diffusion-based models have become a strong alternative for high-quality audio generation. AudioLDM Liu et al. (2023) demonstrated effective text-to-audio synthesis using latent diffusion, while AudioLDM 2 Liu et al. (2024) extended this framework to a unified setting covering speech, music, and general audio. Additional systems, including EZAudio Feng et al. (2024), Make-An-Audio Huang et al. (2023), and Stable Audio Open Evans et al. (2025), further improve scalability and controllability in general audio generation, and speech-focused diffusion models such as VoiceLDM Lee et al. (2024) similarly demonstrate strong performance in text-to-speech tasks. Despite these advances, such models are designed for broad audio domains and do not explicitly capture the temporal structure, frequency characteristics, or clinical relevance of respiratory sounds. In parallel, respiratory-specific research has largely focused on representation learning and task-driven synthesis rather than standalone text-to-audio generation. OPERA Zhang et al. (2024) introduced a foundation-model framework for respiratory audio, emphasizing the importance of curated datasets and task-specific evaluation, while more recent work such as RespAgent Zhang et al. (2026)

proposed a multimodal system that combines diagnosis and synthesis in a closed-loop framework. However, RespAgent does not operate as a standalone text-to-audio model. Its generation relies on both diagnostic context and reference audio representations derived from BEATs tokens, rather than text alone. In practice, this means that synthesis is guided by both prompts and reference audio. In contrast, LungTTA is designed to generate respiratory sounds directly from text prompts, covering cough, breathing, and phonation signals without requiring reference audio. This difference places LungTTA alongside general text-to-audio models, while remaining focused on respiratory data. A comparison of these approaches is shown in Table 1.

Table 1: Positioning of LungTTA relative to general audio generation models, respiratory-specific systems, and augmentation methods.

| Method | Training Data | Backbone | Task | Resp.-Specific | Text-to-Audio |
|---|---|---|---|---|---|
| *General Audio Generation Models* | | | | | |
| AudioLDM (Liu et al., 2023) | General audio | Latent diffusion | Text-to-audio | ✗ | ✓ |
| VoiceLDM (Lee et al., 2024) | General audio (speech) | Latent diffusion | Text-to-speech | ✗ | ✓ |
| AudioLDM 2 (Liu et al., 2024) | General audio | Latent diffusion | Text-to-audio | ✗ | ✓ |
| EZAudio (Feng et al., 2024) | General audio | Diffusion Transformer | Text-to-audio | ✗ | ✓ |
| Make-An-Audio (Huang et al., 2023) | General audio | Diffusion | Text-to-audio | ✗ | ✓ |
| Stable Audio Open (Evans et al., 2025) | General audio | Latent diffusion | Text-to-audio | ✗ | ✓ |
| *Respiratory-Specific Systems* | | | | | |
| OPERA (Zhang et al., 2024) | Respiratory | Foundation model | Representation learning | ✓ | ✗ |
| RespAgent (Zhang et al., 2026) | Respiratory | Flow matching + agent | Diagnosis + synthesis | ✓ | ✗ |
| *Respiratory Data Augmentation* | | | | | |
| No augmentation (Xia et al., 2022) | Respiratory | - | Training only | ✓ | ✗ |
| Classical augmentation (Xia et al., 2022) | Respiratory | Signal transforms | Data augmentation | ✓ | ✗ |
| GAN-based synthesis (Ramesh et al., 2020) | Respiratory | GAN | Synthetic generation | ✓ | ✗ |
| **LungTTA (ours)** | Respiratory | Diffusion Transformer | Text-to-audio | ✓ | ✓ |

## 3 LungTTA

LungTTA is a text-to-audio (TTA) framework for synthesizing high-fidelity respiratory sounds, fine-tuned on publicly available recordings to specialize in cough, breathing, and phonation. As illustrated in Figure 1, the pipeline integrates a variational autoencoder (VAE), a T5-based text conditioning module, a transformer-based diffusion backbone with cross-attention and a noise scheduler, and a retrieval memory. The latent diffusion backbone is adapted from Stable Audio Open (Evans et al., 2025), where audio is compressed into a latent representation using a VAE, generated in latent space via a diffusion transformer (DiT), and decoded back into waveform space. During training, real respiratory audio is passed through the VAE encoder to guide latent alignment, while inference is performed purely from text and conditioning inputs. On top of this backbone, LungTTA introduces respiratory-specific prompt conditioning, retrieval-guided memory augmentation, and watermarking for traceability.

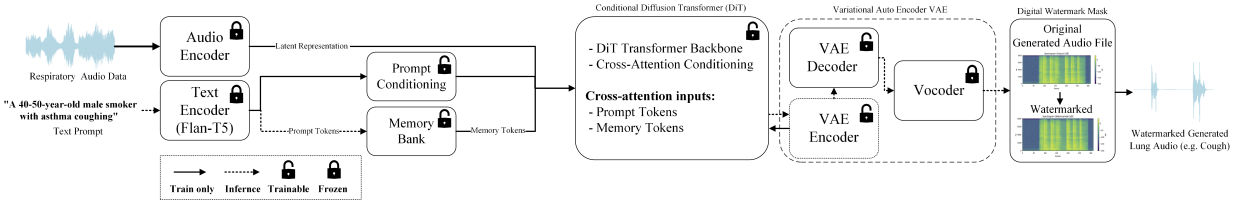



Figure 1: Overview of the proposed LungTTA pipeline. The framework takes textual respiratory prompts as input and generates synthetic lung sounds through a latent diffusion architecture.

Text conditioning is provided through a T5-based encoder. During fine-tuning, the pretrained backbone is exposed to prompts derived from dataset metadata, allowing it to shift from general-purpose audio generation toward clinically meaningful respiratory events. The LungTTA generated audio samples are available at `https://lungtta.github.io/audio_samples/`, and the source code will be released upon acceptance.

### 3.1 Architectural Components

**Backbone and Fine-Tuning** LungTTA builds upon the Stable Audio Open architecture (Evans et al., 2025), which consists of a variational audio autoencoder (156M parameters), a T5-based text encoder (109M parameters), and a diffusion transformer (DiT) with approximately 1.06B parameters operating in latent space. The autoencoder compresses raw waveforms into a 64-channel latent representation at a reduced temporal resolution (∼21.5 Hz), which enables efficient generation of high-resolution audio at 44.1 kHz. Rather than training a model from scratch, we fine-tune this pretrained backbone on respiratory audio, allowing it to retain general audio generation capabilities while adapting to domain-specific acoustic patterns. In contrast to the original backbone (∼1.3B total parameters), the LungTTA variant used here contains 96.41M *trainable* parameters, making it feasible to train on moderate GPU resources. Training follows a latent diffusion objective, where the model learns to denoise latent representations conditioned on text prompts.

**Retrieval-Based Memory Bank** LungTTA includes a retrieval-based memory component. We define a set of conditioning embeddings $\{e_i\}_{i=1}^{N}$ extracted from the training data, stored in a latent space of dimension $d$. For a given query embedding $q \in \mathbb{R}^d$, the model retrieves the $k$ nearest neighbors

$$\{e_{i_1}, \ldots, e_{i_k}\} = \underset{\{i_1,\ldots,i_k\}}{\arg\min}\, d(q, e_i), \tag{1}$$

under a chosen distance function $d(\cdot, \cdot)$. These retrieved embeddings are then used as an additional conditioning signal during generation. In our setting, the data distribution is uneven, with common breathing and cough patterns appearing much more frequently than rarer cases. Without additional guidance, the model tends to reproduce these dominant patterns. The retrieval step helps counter this by grounding generation in nearby examples from the dataset, which improves coverage of less frequent acoustic structures and reduces unrealistic outputs.

**Prompt-Based Conditioning** Generation is driven by structured text prompts. Let $p$ denote a prompt describing the target sound, including attributes such as sound type, demographics, smoking status, or recording conditions. The prompt is mapped to a latent embedding

$$H = f_{\text{text}}(p), \tag{2}$$

where $f_{\text{text}}(\cdot)$ is the text encoder. This embedding serves as the main conditioning signal. We then concatenate it with retrieval-based embeddings $Z_{\text{mem}}$ and metadata features $Z_{\text{meta}}$:

$$Z_{\text{cond}} = [H\,;\, Z_{\text{mem}}\,;\, Z_{\text{meta}}], \tag{3}$$

which is passed to the diffusion model. This setup lets the model follow the prompt while still being influenced by examples from the data.

**Prompt Engineering** The prompts are constructed directly from dataset metadata rather than written manually. Each one follows a simple template, for example: *"A person is [activity], recorded in [condition]."* The activity specifies the sound type, such as coughing or breathing, while the condition includes attributes like age, gender, smoking status, or recording setup. This keeps the prompts consistent with the underlying data. For instance, we use prompts such as `"A person is vocalizing the sustained vowel sound /a/."` or `"A person is coughing, male, aged 40-50, smoker, asthmatic."`. Similar strategies have been used in AudioLDM2 Liu et al. (2023) and RespAgent Zhang et al. (2026), where text descriptions are derived from metadata rather than free-form annotation.

**Ethical Considerations and Traceability** Synthetic respiratory audio can be useful, but it also raises practical concerns. Generated samples may resemble real recordings and could be mistaken for genuine data if not clearly identified. To address this, we add a low-amplitude watermark $w(t)$ to each generated waveform:

$$\hat{x}(t) = x(t) + \lambda w(t), \tag{4}$$

with $\lambda \ll 1$ so that the perceptual quality is not affected. This allows synthetic samples to be traced if needed. It is worth noting that this does not prevent removal of the watermark, and it does not ensure clinical validity. Since the model is trained on public datasets, it may also reflect biases present in those sources, including imbalances in demographics or recording conditions. For this reason, generated audio should not be treated as diagnostic evidence. We use it as a supporting tool for data augmentation, and synthetic samples are kept separate from real recordings during evaluation. In addition, we view traceability, transparent reporting, and clear use restrictions as essential safeguards for the responsible development of synthetic respiratory audio.

### 3.2 PRISM: A Domain-Aware Consistency Metric for Respiratory Audio

Conventional metrics such as FAD, KL divergence, and Inception Score measure global distributional similarity between real and generated audio. However, respiratory sounds are often interpreted at the event level, where clinical relevance depends on temporal evolution, breathing-cycle structure, tonal wheeze-like content, transient crackle-like behavior, and the distribution of energy across respiratory frequency bands. To address this limitation, we introduce **PRISM (Pulmonary Respiratory Integrity & Similarity Metric)**, a domain-aware similarity metric defined for a real waveform $x$ and a generated waveform $\hat{x}$ as

$$\mathrm{PRISM}(x, \hat{x}) = \sum_{i=1}^{5} w_i S_i(x, \hat{x}), \tag{5}$$

where $S_i$ are respiratory structure similarity terms and $w_i \geq 0$ are non-negative weights. In our implementation, PRISM combines five interpretable components:

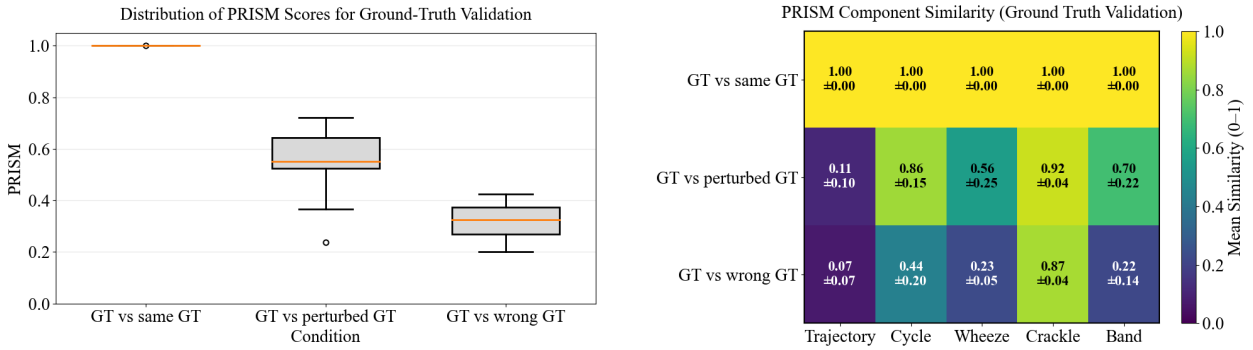
$$\mathrm{PRISM}(x, \hat{x}) = 0.30\, S_{\mathrm{traj}} + 0.20\, S_{\mathrm{cycle}} + 0.20\, S_{\mathrm{wheeze}} + 0.15\, S_{\mathrm{crackle}} + 0.15\, S_{\mathrm{band}}. \tag{6}$$

These components are selected to capture complementary aspects of respiratory sound structure, including temporal evolution, coarse breathing dynamics, tonal coherence, transient events, and frequency energy distribution, which are commonly used in respiratory sound analysis (Pramono et al., 2017). Here, $S_{\mathrm{traj}}$ measures trajectory similarity, defined as the agreement in the temporal evolution of spectral features over time using Mel-frequency cepstral coefficients (MFCC) aligned via dynamic time warping (DTW), $S_{\mathrm{cycle}}$ measures similarity of the root mean square (RMS) energy envelope and coarse respiratory cycle balance, $S_{\mathrm{wheeze}}$ measures band-limited phase coherence in the 400-1600 Hz range, $S_{\mathrm{crackle}}$ measures similarity of frame-wise kurtosis as a proxy for transient crackle-like structure, and $S_{\mathrm{band}}$ measures band-energy similarity, defined as the agreement in the distribution of spectral energy across predefined respiratory frequency bands. Higher PRISM values indicate stronger agreement between generated and real respiratory structure.

**Metric Validation on Ground-Truth Data** We test PRISM using 20 recordings from the ground-truth dataset under three pairing settings: identical pairs (GT vs same GT), perturbed pairs (GT vs slightly modified GT), and mismatched pairs (GT vs different GT). As expected, identical pairs produce the highest scores, perturbed pairs fall in between, and mismatched pairs give the lowest values. This trend is shown in Figure 2. The behavior indicates that PRISM responds to both signal distortion and structural differences in the audio, rather than only overall similarity.

## 4 Experiments

**Dataset** We train and evaluate LungTTA using nine publicly available respiratory audio datasets, comprising a total of 116,660 recordings covering coughs, breathing patterns, and vowel phonation. Table 2 provides an overview of the datasets, including recording devices and sample counts. As the datasets originate from different sources, a preprocessing step was required to standardize them. All audio was resampled to 16 kHz, corrupted files were removed, and filenames were normalized into a consistent format. Metadata was retained and converted into a unified JSON representation to support prompt conditioning and downstream evaluation. The data was split into training (80%), validation (10%), and testing (10%) sets. To




(a) Distribution of overall PRISM scores across controlled pairing conditions.

(b) Component-wise PRISM similarity (mean ± standard deviation across sample pairs).

Figure 2: PRISM validation on 20 ground-truth recordings under controlled conditions. Left, identical pairs achieve the highest scores, with decreasing values under perturbation and mismatch. Right, component-wise similarity shows consistent trends across trajectory, cycle, wheeze, crackle, and band features.

support conditioning and analysis, recordings were grouped into 13 categories based on their original labels and recording protocols, including cough, breathing, vowel phonation, counting tasks, and stethoscope-based recordings. Where available, finer-grained labels such as *deep* and *shallow* breathing, or variations in cough intensity (e.g., *heavy* or *shallow*), were retained from the source annotations.

Table 2: Summary of datasets and number of recordings.

| **Dataset** | **Device** | **# Recordings.** |
|---|---|---|
| COVID-19 Sounds (Han et al., 2022) | Microphone | 53,449 |
| UK COVID-19 (Coppock et al., 2024) | Microphone | 25,706 |
| COUGHVID (Orlandic et al., 2021) | Microphone | 20,072 |
| CoronaHack (Thandu & Gera, 2024) | Microphone | 1,400 |
| MMLung (Mosuily et al., 2023) | Microphone | 560 |
| Vowels (David Andres Rubiano Venegas, 2019) | Microphone | 1,676 |
| HF Lung (Hsu et al., 2022) | Stethoscope | 9,765 |
| Respiratory TR (Altan et al., 2017) | Stethoscope | 3,696 |
| KAUH Lung (Fraiwan et al., 2021) | Stethoscope | 336 |

**Training** LungTTA is trained as a text-conditioned latent diffusion model for synthetic respiratory audio generation on a high-performance computing node equipped with a single NVIDIA A100 GPU, 48 CPU cores and 180 GB RAM, using PyTorch and Librosa (McFee et al., 2015). The dataset contains 116,660 recordings, split into 93,328 training, 11,666 validation, and 11,666 test samples. Text prompts are encoded using a T5 encoder, while audio is mapped into a latent space using a pretrained VAE that remains frozen during training. The generative backbone is a diffusion transformer (DiT) with 24 layers, 1536 embedding dimension and 24 attention heads, operating on 64-channel latent representations. In addition to text conditioning, temporal variables (start time and total duration) are included as global conditioning signals. A retrieval-based memory module is used during training, where the top-$k$ nearest embeddings ($k = 4$) from a precomputed memory bank are incorporated as auxiliary conditioning, where $k = 4$ balances diversity and conditioning stability based on empirical observations and prior retrieval-augmented methods (Lewis et al., 2020). The model is optimized using AdamW with a learning rate of $5 \times 10^{-5}$ and weight decay $1 \times 10^{-3}$, and trained for 250,000 steps with checkpoints saved every 5,000 steps and validation performed at each epoch. Classifier-free guidance is applied during sampling with guidance scales between 4 and 7. The full model contains approximately 96.41 M parameters and requires around 30 hours of training.

## 5 Results

We evaluate LungTTA using a mix of objective metrics, listening studies, ablation experiments, and a downstream classification task. The aim is to understand both the quality of the generated audio and how well it follows the conditioning prompts, as well as whether the synthetic data is useful in practice. A qualitative example is shown in Figure 3, where LungTTA produces spectrograms that better match the timing and energy patterns of real cough signals compared to baseline models. The following sections report quantitative results, prompt diversity experiments, comparisons with existing methods, ablations, and downstream performance.

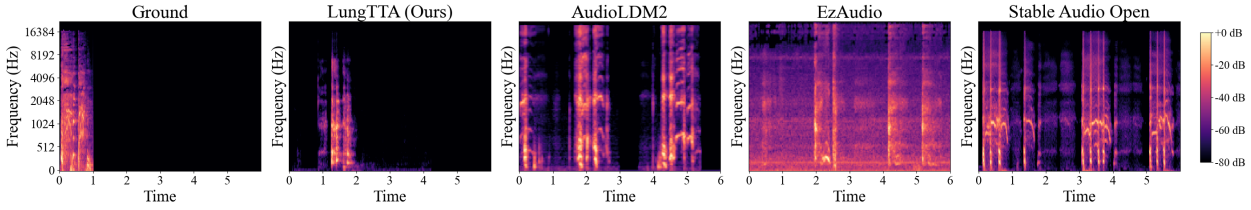



Figure 3: Spectrogram comparison for the prompt *"a cough from a 48-year-old male smoker with no asthma and no COPD"*. LungTTA produces clearer and more localized cough events, with energy concentrated in expected frequency bands. In contrast, baseline models tend to produce smoother or noisier patterns that do not align as well with the ground-truth structure.

### 5.1 Objective Evaluation

We evaluate generated lung sounds using Fréchet Audio Distance (FAD), Kullback-Leibler divergence (KL), Inception Score (IS), and PRISM. Lower values indicate better performance for FAD and KL, while higher values are preferred for IS and PRISM. To study the effect of prompt diversity, we test 1, 5, 10, 20, and 50 conditioning prompts under the same setup. As shown in Figure 4, FAD drops from 17.33 (1 prompt) to 1.81 (50 prompts), and KL from 1.94 to 0.19, while IS increases from 1.12 to 1.33. PRISM stays within a narrower range, between 0.17 and 0.21. With fewer than 5 prompts, the results vary more, likely due to limited coverage of the data distribution. Increasing the number of prompts improves diversity and leads to more stable outputs. After around 20 prompts, the gains become smaller, suggesting that most of the dominant variations are already captured. Based on this, we use 20 prompts in the remaining experiments.

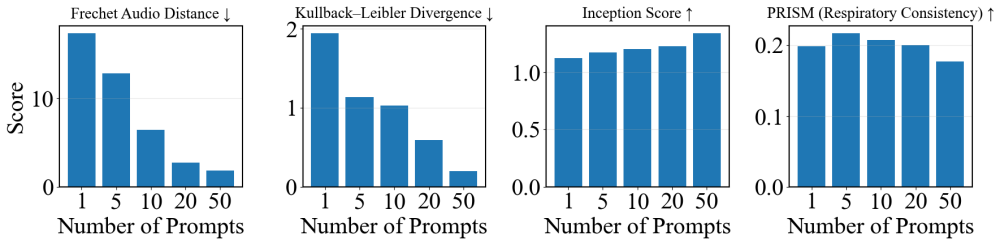



Figure 4: Effect of prompt diversity on LungTTA performance.

#### 5.1.1 Comparison with State-of-the-Art Methods

All models are evaluated under identical conditions. Table 3 shows that LungTTA achieves the lowest FAD (2.72) and KL (0.50). Make-An-Audio achieves the highest IS (1.54), while RespAgent obtains the highest PRISM (0.24). This is expected, as RespAgent conditions on both text and reference audio. LungTTA, which is text-only, achieves a comparable PRISM score of 0.23.

Table 3: Comparison with state-of-the-art text-to-audio models.

| Model | FAD ↓ | KL ↓ | IS ↑ | PRISM ↑ |
|---|---|---|---|---|
| AudioLDM2 (Liu et al., 2023) | 6.21 | 0.71 | 1.35 | 0.15 |
| EzAudio (Feng et al., 2024) | 6.59 | 0.52 | 1.46 | 0.14 |
| Make-An-Audio (Huang et al., 2023) | 15.95 | 1.34 | **1.54** | 0.14 |
| StableAudioOpen (Evans et al., 2025) | 6.73 | 0.67 | 1.39 | 0.21 |
| RespAgent (Zhang et al., 2026) | 8.53 | 0.53 | 1.44 | **0.24** |
| **LungTTA (Ours)** | **2.72** | **0.50** | 1.22 | 0.23 |

#### 5.1.2 Ablation Study

We analyze the contribution of the memory bank and watermarking using FAD, KL, IS, and PRISM (Table 4). The baseline model achieves an FAD of 6.73 and a PRISM score of 0.21. Adding watermarking produces minimal change in FAD and IS, although KL decreases from 0.67 to 0.47. IS remains largely unchanged (1.39 to 1.38), indicating that watermarking does not affect generation quality. Introducing the memory bank leads to a substantial improvement in FAD (6.73 to 2.72), while IS decreases from 1.39 to 1.23. The full LungTTA model achieves a KL of 0.50, while the lowest KL (0.47) is observed when watermarking is applied alone. The full model also achieves the highest PRISM score (0.23). Overall, the memory component provides the main improvement in FAD, while watermarking has limited effect on FAD and IS but reduces KL.

Table 4: Ablation study evaluating the impact of the proposed memory and watermarking modules.

| Setting | Memory Bank | Watermarking | FAD ↓ | KL ↓ | IS ↑ | PRISM ↑ |
|---|---|---|---|---|---|---|
| Baseline Model | × | × | 6.73 | 0.67 | **1.39** | 0.21 |
| Baseline + Watermarking | × | ✓ | 6.74 | **0.47** | 1.38 | 0.20 |
| Baseline + Memory Bank | ✓ | × | 2.72 | 0.59 | 1.23 | 0.18 |
| **LungTTA (Full Model)** | ✓ | ✓ | **2.72** | **0.50** | 1.22 | **0.23** |

#### 5.1.3 Downstream Evaluation

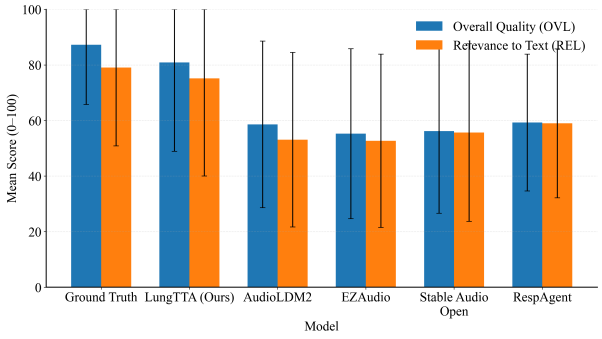
To assess the usefulness of the generated data for downstream tasks, we follow a realistic evaluation protocol inspired by prior work on audio-based COVID-19 detection (Han et al., 2022). In particular, we adopt a participant-independent data split and consistent training and evaluation settings to avoid over-optimistic performance estimates. For the baseline methods, we use a VGGish-based architecture, where VGGish, a VGG-like Convolutional neural network pretrained on AudioSet, serves as the feature extraction backbone, followed by pooling and fully connected layers for classification. For LungTTA, we additionally evaluate performance using an Audio Spectrogram Transformer (AST)-based model (Gong et al., 2021), where mid-level representations are extracted from the transformer backbone and used for classification. All models are trained and evaluated on identical data splits, and the decision threshold is selected based on validation performance before testing. For training, the real dataset is imbalanced, consisting of 7,660 positive and 14,188 negative samples. To mitigate this imbalance and improve generalization, we augment the positive class with 7,000 synthetic samples generated by LungTTA, resulting in a more balanced training set with 14,660 positive and 14,188 negative samples. As shown in Table 5, LungTTA consistently improves downstream classification performance over classical and GAN-based augmentation. The VGGish-based model achieves an F1 score of 0.6213, compared to 0.6087 with classical augmentation and 0.5801 without augmentation. The AST-based configuration achieves an AUC of 0.7791 and sensitivity of 0.7027, compared to 0.7701 AUC and 0.6995 sensitivity for the VGGish-based LungTTA model.

Table 5: Downstream COVID-19 cough classification performance under identical evaluation settings.

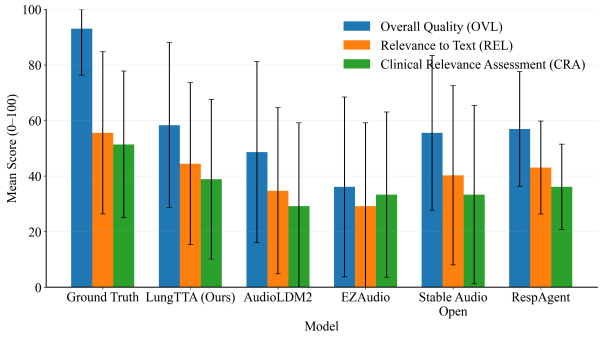
| Method | Model | F1 ↑ | AUC ↑ | Sensitivity ↑ | Specificity ↑ |
|---|---|---|---|---|---|
| No Augmentation | VGGish-based | 0.5801 | 0.7331 | 0.6985 | 0.6397 |
| Classic Augmentation | VGGish-based | 0.6087 | 0.7631 | 0.6672 | 0.7333 |
| GAN | VGGish-based | 0.5029 | 0.7540 | 0.4123 | **0.8846** |
| **LungTTA (Ours)** | VGGish-based | **0.6213** | 0.7701 | 0.6995 | **0.7196** |
| **LungTTA (Ours)** | AST (mid-layer) | 0.6170 | **0.7791** | **0.7027** | 0.7080 |

### 5.2 Subjective Evaluation

We complement the objective metrics with human ratings to assess audio quality, alignment with the input prompt, and clinical usefulness. In the general public study ($N = 32$), we evaluated 6 recordings per model (192 ratings per model), where each sample was rated on Overall Quality (OVL) and Relevance to Text (REL) using a 5-point Likert scale mapped to a 0-100 range. Ground truth achieved 87.23 (OVL) and 79.03 (REL), while LungTTA achieved 80.91 (OVL) and 75.13 (REL). Among the synthetic models, RespAgent obtained 59.27 (OVL) and 58.97 (REL), followed by AudioLDM2 with 58.60 (OVL) and 53.09 (REL), Stable Audio Open with 56.18 (OVL) and 55.65 (REL), and EZAudio with 55.24 (OVL) and 52.69 (REL), showing that LungTTA consistently outperforms all other models by more than 20 points in both metrics. For clinical usefulness, we conducted an expert evaluation with three respiratory-health professionals using the same recordings (18 ratings per model), where in addition to OVL and REL, Clinical Relevance for Assessment (CRA), a metric developed in consultation with clinical experts to assess whether an audio sample contains meaningful respiratory patterns that may support clinical interpretation, was rated on a 0-100 scale. LungTTA achieved 58.33 (OVL), 44.44 (REL), and 38.89 (CRA), compared to RespAgent with 56.94 (OVL), 43.06 (REL), and 36.11 (CRA), AudioLDM2 with 48.61 (OVL), 34.72 (REL), and 29.17 (CRA), Stable Audio Open with 55.56 (OVL), 40.28 (REL), and 33.33 (CRA), and EZAudio with 36.11 (OVL), 29.17 (REL), and 33.33 (CRA), indicating that LungTTA achieves the highest scores across all expert-rated metrics while maintaining consistent improvements over each comparison model.




(a) General public evaluation results ($N = 32$). Error bars indicate standard deviation.




(b) Expert evaluation across OVL, REL, and CRA (0-100 scale, mean ± standard deviation). Scores reflect ratings from three respiratory-health professionals.

Figure 5: Subjective evaluation of generated respiratory audio. Left, ratings from 32 general listeners for Overall Quality (OVL) and Relevance to Text (REL). Right, ratings from three respiratory-health professionals for OVL, REL, and Clinical Relevance for Assessment (CRA). Scores are shown on a 0-100 scale, where 100 is the highest possible score.

## 6 Discussion

The results show that LungTTA improves how closely generated audio matches real respiratory recordings, without being the best on every metric. It achieves the lowest FAD and KL (2.72 and 0.50), compared with AudioLDM2 (6.21, 0.71), EZAudio (6.59, 0.52), Stable Audio Open (6.73, 0.67), and RespAgent (8.53, 0.53), suggesting better distributional alignment. It does not reach the highest IS (1.22 vs 1.54 for Make-An-Audio) and remains slightly below RespAgent on PRISM (0.23 vs 0.24), likely because RespAgent uses reference audio while LungTTA is text-only. The ablation study shows that most of the FAD improvement comes from the memory component (6.73 to 2.72), while the full model achieves the highest PRISM (0.23) and watermarking has little effect on perceptual quality. The listening study supports this, with LungTTA achieving 80.91 (OVL) and 75.13 (REL), compared with around 55–59 for other models, and similar improvements observed in expert-rated OVL, REL, and CRA. In the downstream task, LungTTA improves AUC from 0.7331 (no augmentation) and 0.7631 (classical augmentation) to 0.7701, with the highest AUC of 0.7791 achieved using AST. Overall, these results indicate that LungTTA generates useful task-relevant samples rather than simple augmentation noise, although improvements in PRISM are smaller than for distributional metrics and evaluation remains limited to a single downstream task.

## 7 Conclusion

This work introduced LungTTA, a prompt-based text-to-audio framework for generating respiratory sounds using latent diffusion models. LungTTA enables controllable synthesis of cough, breathing, and phonation signals from structured prompts, while incorporating retrieval-based memory guidance and watermarking for traceability. The results show that LungTTA improves the realism and usefulness of synthetic respiratory audio compared to existing baselines, and can support downstream learning in data-scarce settings. LungTTA is not intended to replace real data collection, but to act as a complementary tool for augmenting limited datasets and supporting model development. By enabling scalable and reproducible generation of respiratory audio, it provides a practical pathway for improving robustness and coverage in respiratory health applications. Future work will extend evaluation across more diverse clinical conditions, recording environments, and downstream tasks, and further improve prompt fidelity and clinically relevant structure in generated audio.

## 8 Compliance with Ethical Standards

This work uses publicly available respiratory sound datasets in line with their licenses and data sharing agreements. All human data collection was approved by relevant ethics committees and conducted according to standard ethical and data protection guidelines. Written consent was obtained from all participants.

## References

Gokhan Altan, Yakup Kutlu, Yusuf Garbi, Adnan Ozhan Pekmezci, and Serkan Nural. Multimedia Respiratory Database (RespiratoryDatabase@TR): Auscultation Sounds and Chest X-rays. *Natural and Engineering Sciences*, 2(3):59–72, 2017. doi: 10.28978/nesciences.349282.

Tanujit Chakraborty, Ujjwal Reddy K S, Shraddha M Naik, Madhurima Panja, and Bayapureddy Manvitha. Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. *Machine Learning: Science and Technology*, 5(1):11001, 1 2024. doi: 10.1088/2632-2153/ad1f77. URL `https://dx.doi.org/10.1088/2632-2153/ad1f77`.

Harry Coppock et al. Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers. *Nature Machine Intelligence*, 6(2):229–242, 2 2024. ISSN 25225839. doi: 10.1038/s42256-023-00773-8.

David Andres Rubiano Venegas. Dataset of Vowels, 2019. URL `https://www.kaggle.com/datasets/darubiano57/dataset-of-vowels/data`.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888461.

Tiantian Feng, Dimitrios Dimitriadis, and Shrikanth S. Narayanan. Can Synthetic Audio From Generative Foundation Models Assist Audio Recognition and Speech Modeling? In *International Speech Communication Association*, pp. 542–546, 9 2024. doi: 10.21437/interspeech.2024-1350.

Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913, 2021. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2021.106913. URL `https://www.sciencedirect.com/science/article/pii/S2352340921001979`.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, Pietro Cicuta, and Cecilia Mascolo. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *npj Digital Medicine*, 5(1), 12 2022. ISSN 23986352. doi: 10.1038/s41746-021-00553-x.

Fu Shun Hsu, Shang Ran Huang, Chien Wen Huang, Yuan Ren Cheng, Chun Chieh Chen, Jack Hsiao, Chung Wei Chen, and Feipei Lai. A Progressively Expanded Database for Automated Lung Sound Analysis: An Update. *Applied Sciences (Switzerland)*, 12(15), 8 2022. ISSN 20763417. doi: 10.3390/app12157623.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/huang23i.html`.

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. VoiceLDM: Text-to-Speech with Environmental Context. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12566–12571, 2024. doi: 10.1109/ICASSP48485.2024.10448268.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Xiaochen Li, Xiaopei Cao, Mingzhou Guo, Min Xie, and Xiansheng Liu. Trends and risk factors of mortality and disability adjusted life years for chronic respiratory diseases from 1990 to 2017: systematic analysis for the Global Burden of Disease Study 2017. *BMJ*, 368, 2020. doi: 10.1136/bmj.m237. URL `https://www.bmj.com/content/368/bmj.m237`.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *ICML International Conference on Machine Learning*, pp. 21450 – 21474, 1 2023. URL `http://arxiv.org/abs/2301.12503`.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 564–572, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681688. URL `https://doi.org/10.1145/3664647.3681688`.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. *PROC. OF THE 14th PYTHON IN SCIENCE CONF. (SCIPY*, 2015. URL https://github.com/bmcfee/librosa.

Mohammed Mosuily, Lindsay Welch, and Jagmohan Chauhan. MMLung: Moving Closer to Practical Lung Health Estimation using Smartphones. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023-August, pp. 2333–2337. International Speech Communication Association, 2023. doi: 10.21437/Interspeech.2023-721.

Ebrahim Nemati, Xuhai Xu, Viswam Nathan, Korosh Vatanparvar, Tousif Ahmed, Md. Mahbubur Rahman, Dan McCaffrey, Jilong Kuang, and Alex Gao. Ubilung: Multi-Modal Passive-Based Lung Health Assessment. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 551–555, 2022. doi: 10.1109/ICASSP43922.2022.9746614.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. Towards pre-training an effective respiratory audio foundation model. In *Proceedings of Interspeech*, Rotterdam, The Netherlands, 2025. ISCA.

Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1), 12 2021. ISSN 20524463. doi: 10.1038/s41597-021-00937-4.

Renard Xaviero Adhi Pramono, Stuart Bowyer, and Esther Rodriguez-Villegas. Automatic adventitious respiratory sound analysis: A systematic review. *PLoS One*, 12(5):e0177926, 2017. doi: 10.1371/journal.pone.0177926.

Vishwajith Ramesh, Korosh Vatanparvar, Ebrahim Nemati, Viswam Nathan, Md Mahbubur Rahman, and Jilong Kuang. CoughGAN: Generating Synthetic Coughs that Improve Respiratory Disease Classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 5682–5688, 2020. doi: 10.1109/EMBC44109.2020.9175597.

Joan B. Soriano and et al. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Respiratory Medicine*, 5(9):691–706, 9 2017. ISSN 22132619. doi: 10.1016/S2213-2600(17)30293-X.

Asha Latha Thandu and Pradeepini Gera. CoronaHack-Respiratory-Sound-Dataset, 2024. URL https://dx.doi.org/10.21227/z6eq-hw49.

Tong Xia, Jing Han, and Cecilia Mascolo. Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine*, 247(22):2053–2061, 2022. doi: 10.1177/15353702221115428. URL https://doi.org/10.1177/15353702221115428.

Pengfei Zhang, Tianxin Xie, Minghao Yang, and Li Liu. Resp-agent: An agent-based system for multimodal respiratory sound generation and disease diagnosis. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=ZkoojtEm3W.

Yuwei Zhang, Tong Xia, Jing Han, Yu Y Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 27024–27055, 2024.

# Appendix

This appendix provides additional technical and experimental details supporting the main paper. It includes formal definitions of the objective and subjective evaluation metrics, details of the expert clinical assessment protocol, and the downstream classification setup. We also report exploratory architectural variants results to contextualize the final design choices. In addition, the appendix contains the conditioning prompts used for generation, representative prompt variations, and a reconstruction of the survey interface used in the listening study for both general participants and clinical experts.

## A Evaluation Metrics

We evaluate LungTTA using a combination of objective statistical metrics and subjective perceptual assessments in order to capture realism, distributional similarity, and clinical usefulness in the generated respiratory sounds. All objective metrics are computed within a unified evaluation pipeline in which feature embeddings are extracted from generated samples and their corresponding real recordings before metric computation.

### A.1 Objective Metrics

#### A.1.1 Fréchet Audio Distance (FAD)

Fréchet Audio Distance quantifies statistical similarity between real and synthetic audio distributions in the embedding space induced by the `FrechetAudioDistance` backend. This representation is intended to retain perceptually relevant attributes such as spectral texture and temporal structure, making FAD a useful proxy for perceptual realism in respiratory audio. Let $\boldsymbol{\mu}_r$ and $\boldsymbol{\mu}_g$ denote the mean feature embeddings of real and generated samples, and let $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_g$ denote their covariance matrices. FAD is defined as

$$\mathrm{FAD} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\sqrt{\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g}\right). \tag{A.1}$$

Lower values indicate that synthetic sounds align more closely with the perceptual distribution of real lung recordings, with zero representing a perfect match.

#### A.1.2 Kullback–Leibler Divergence (KL)

To assess alignment between the probability distributions of real and generated embeddings, we compute the Kullback–Leibler divergence in both sigmoid and softmax activation domains, as returned by the evaluation pipeline. KL divergence measures how much the generated distribution deviates from the real reference distribution and is expressed as

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{A.2}$$

Lower values indicate closer distributional agreement, with values approaching zero representing near-identical distributions.

#### A.1.3 Inception Score (IS)

We also report Inception Score to evaluate the joint quality and diversity of generated audio. IS is computed from classifier outputs over multiple splits of the synthetic dataset and is defined as

$$\mathrm{IS} = \exp\Big(\mathbb{E}_x \, D_{\mathrm{KL}}\left(p(y|x) \,\|\, p(y)\right)\Big), \tag{A.3}$$

where $p(y|x)$ is the conditional class distribution for audio sample $x$, and $p(y)$ is the marginal distribution across samples. Higher scores indicate more confident and diverse generations.

### A.2 Subjective Evaluation

Participants rated each audio clip on Overall Quality (OVL) and Relevance to Text (REL) using a 0–100 scale. For each clip $n$, scores are averaged across $K$ participants,

$$\mathrm{OVL}_n = \frac{1}{K}\sum_{k=1}^{K} o_{n,k}, \qquad \mathrm{REL}_n = \frac{1}{K}\sum_{k=1}^{K} r_{n,k}. \tag{A.4}$$

### A.3 Expert Clinical Assessment

An expert evaluation was conducted with three respiratory-health professionals. Each clip was rated on Overall Quality (OVL), Relevance to Text (REL), and Clinical Relevance for Assessment (CRA). For clip $n$, the CRA score is computed as

$$\mathrm{CRA}_n = \frac{1}{K}\sum_{k=1}^{K} c_{n,k}, \tag{A.5}$$

where $K = 3$. Model-level scores are obtained by averaging across all evaluated clips,

$$\mathrm{CRA}_{\mathrm{model}} = \frac{1}{KN}\sum_{n=1}^{N}\sum_{k=1}^{K} c_{n,k}. \tag{A.6}$$

### A.4 Downstream Evaluation

We report standard classification metrics. For completeness, the definitions are given below.

**F1 Score**

$$\mathrm{F1} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{A.7}$$

**Sensitivity**

$$\mathrm{Sensitivity} = \frac{TP}{TP + FN} \tag{A.8}$$

**Specificity**

$$\mathrm{Specificity} = \frac{TN}{TN + FP} \tag{A.9}$$

**AUC**

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(t)\, d(\mathrm{FPR}(t)) \tag{A.10}$$

## B Exploratory Architectural Variants and Additional Experimental Results

To better understand the design space of text-to-audio generation for respiratory sounds, we conducted exploratory experiments evaluating alternative architectural modifications beyond the proposed memory and watermarking framework. These include squeeze-and-excitation (SE) encoders and channel-wise gating mechanisms. Table B.1 summarizes the performance of these variants. The results show that these modifications do not provide consistent improvements in FAD and KL, indicating limited gains in aligning generated audio with real respiratory data. For example, adding SE encoding increases FAD from 6.73 to 7.96, while combining SE with channel-wise gating results in further shifts in performance (FAD = 9.19, KL = 6.18). We also evaluate removing these components from the proposed model. The results indicate that these configurations do not lead to improvements over the full model. Overall, these findings suggest that architectural techniques commonly used in general audio modeling may not directly translate to respiratory sound generation, while memory-based conditioning provides a more stable and effective solution.

Table B.1: Exploratory experiments evaluating alternative architectural modifications. These configurations do not consistently improve performance compared to the proposed memory and watermarking design.

| Configuration | FAD ↓ | KL ↓ | IS ↑ | PRISM ↑ |
|---|---|---|---|---|
| **Baseline (No Memory, No Watermarking)** | 6.73 | 0.67 | **1.39** | 0.21 |
| **Memory + Watermarking (Proposed)** | **2.72** | **0.19** | 1.22 | **0.23** |
| *Additional Modifications* | | | | |
| + SE Encoder | 7.96 | 5.92 | 1.03 | 0.14 |
| + Channel-Wise Gating | 8.86 | 7.42 | 0.55 | 0.13 |
| + SE Encoder + Channel-Wise Gating | 9.19 | 6.18 | 0.58 | 0.13 |
| *Component Variants* | | | | |
| − SE Encoder | 9.71 | 8.33 | 0.84 | 0.13 |
| − Channel-Wise Gating | 9.73 | 6.53 | 0.57 | 0.20 |

# C Downstream Classification Setup and Design

We evaluate the generated data using two downstream pipelines: a VGGish-based model operating on log-Mel spectrogram patches with attention pooling, and an Audio Spectrogram Transformer (AST) model that captures longer-range temporal dependencies. Both are trained and evaluated under consistent settings to assess the impact of synthetic data across different modeling approaches.

## C.1 VGGish-Based Downstream Model

To support the downstream results reported in Table 5, we provide the detailed training and evaluation setup of the VGGish-based classifier (Han et al., 2022). The model operates on log-Mel spectrogram patches extracted from cough recordings and aggregates patch-level features using attention-based pooling. Each waveform $x$ is resampled to 16 kHz, converted to mono, and normalized to a fixed duration. A log-Mel spectrogram is computed and scaled to the range $[-1, 1]$:

$$\mathbf{S} = \mathrm{MelSpec}(x), \quad \tilde{\mathbf{S}} = 2 \cdot \frac{\mathbf{S} - \min(\mathbf{S})}{\max(\mathbf{S}) - \min(\mathbf{S})} - 1. \tag{C.1}$$

This transformation ensures a consistent dynamic range across all samples.

The spectrogram is divided into fixed-width patches:

$$\tilde{\mathbf{S}} \to \{\mathbf{P}_1, \ldots, \mathbf{P}_T\}, \tag{C.2}$$

where each patch captures a local time–frequency segment of the cough signal.

Each patch is encoded using the VGGish backbone:

$$\mathbf{h}_t = f_{\mathrm{VGGish}}(\mathbf{P}_t), \quad \mathbf{h}_t \in \mathbb{R}^{512}. \tag{C.3}$$

These embeddings represent local acoustic features such as spectral shape and temporal energy.

Patch-level features are aggregated using attention. The attention score for each patch is computed as

$$a_t = \mathbf{w}_2^\top \tanh(\mathbf{W}_1 \mathbf{h}_t + \mathbf{b}_1), \tag{C.4}$$

and converted into normalized weights

$$\alpha_t = \frac{\exp(a_t)}{\sum_j \exp(a_j)}, \tag{C.5}$$

which indicate the relative importance of each patch.

Table C.1: VGGish-based downstream classification setup.

| **Component** | **Setting** |
|---|---|
| *Model Architecture* | |
| Backbone | VGGish-style CNN |
| Feature dimension | 512 |
| Pooling | Attention-based pooling |
| Classifier | MLP ($512 \to 128 \to 64 \to 1$) |
| Dropout | 0.5 |
| *Input Processing* | |
| Sampling rate | 16 kHz |
| Audio duration | 10 seconds |
| Channels | Mono |
| FFT size | 1024 |
| Window length | 400 samples |
| Hop length | 160 samples |
| Mel bins | 64 |
| Representation | Log-Mel spectrogram (scaled to [-1,1]) |
| Patch width | 96 frames |
| Patch extraction | Non-overlapping (stride = 96) |
| *Training Setup* | |
| Batch size | 32 |
| Epochs | 50 |
| Optimizer | AdamW |
| Learning rate (backbone) | $1 \times 10^{-5}$ |
| Learning rate (head) | $1 \times 10^{-3}$ |
| Weight decay | $1 \times 10^{-3}$ |
| Loss | Binary focal loss ($\alpha = 0.85$, $\gamma = 2.0$) |
| Class balancing | Weighted random sampling (inverse frequency) |
| Scheduler | CosineAnnealingWarmRestarts |
| *Augmentation* | |
| Audio augmentations | Noise, time stretch, pitch shift, temporal shift |
| GAN augmentation | LSGAN patch replacement (p = 0.6, positives only) |
| GAN epochs | 35 |
| Latent dimension | 128 |
| *Evaluation* | |
| Threshold selection | Validation-based (Youden's J) |
| Metrics | F1, AUC, Sensitivity, Specificity |

The final representation is obtained by weighted pooling:

$$\mathbf{c} = \sum_t \alpha_t \mathbf{h}_t. \tag{C.6}$$

This produces a single vector summarizing the entire cough recording.

The classifier predicts the probability of the positive class:

$$\hat{y} = \sigma(f_{\text{cls}}(\mathbf{c})), \tag{C.7}$$

where $\sigma$ is the sigmoid function.

Training uses binary focal loss:

$$\mathcal{L} = -\alpha y(1-\hat{y})^\gamma \log(\hat{y}) - (1-\alpha)(1-y)\hat{y}^\gamma \log(1-\hat{y}), \tag{C.8}$$

which emphasizes hard examples and mitigates class imbalance.

Synthetic augmentation is introduced using a generator:

$$\tilde{\mathbf{P}} = G(\mathbf{z}), \quad \mathbf{z} \sim \mathcal{N}(0, I), \tag{C.9}$$

where generated patches replace real positive patches during training to increase diversity.

The generator and discriminator are trained using least-squares objectives:

$$\mathcal{L}_D = \frac{1}{2}(D(\mathbf{P}) - 1)^2 + \frac{1}{2}D(G(\mathbf{z}))^2, \tag{C.10}$$

$$\mathcal{L}_G = (D(G(\mathbf{z})) - 1)^2. \tag{C.11}$$

These losses stabilize training and improve the quality of generated patches.

During evaluation, the decision threshold is selected from the ROC curve:

$$J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau), \tag{C.12}$$

$$\tau^* = \arg\max_{\tau} J(\tau), \tag{C.13}$$

which maximizes the balance between sensitivity and specificity.

### C.2 Audio Spectrogram Transformer (AST) Model

To evaluate the practical utility of the generated audio, we design a downstream classification task using an Audio Spectrogram Transformer (AST) backbone (Gong et al., 2021). The model extracts mid-layer representations (block 7), which provide a balance between local acoustic features and higher-level temporal semantics. The full setup is summarized in Table C.2.

Each waveform $x$ is resampled to 16 kHz, converted to mono, high-pass filtered, and padded or truncated to a fixed duration before being processed by the AST feature extractor.

We adopt a mid-layer feature extraction strategy. Let $\mathbf{H}^{(l)}$ denote the hidden representation at transformer block $l$. The feature vector is extracted from block 7 using the [CLS] token:

$$\mathbf{h} = \mathbf{H}^{(7)}_{\mathrm{CLS}} = \mathrm{AST}_{\text{block }7}(x)_{\mathrm{CLS}}. \tag{C.14}$$

This representation captures both local spectral structure and longer-range temporal dependencies.

The feature is regularized using dropout:

$$\tilde{\mathbf{h}} = \mathrm{Dropout}(\mathbf{h}, p = 0.65), \tag{C.15}$$

The classifier computes the logit:

$$z = \mathbf{W}\tilde{\mathbf{h}} + b, \tag{C.16}$$

and predicts the probability of the positive class:

$$\hat{y} = \sigma(z), \tag{C.17}$$

Training is performed using binary cross-entropy with logits:

$$\mathcal{L} = -\left[y \log \sigma(z) + (1 - y) \log\left(1 - \sigma(z)\right)\right]. \tag{C.18}$$

To improve robustness, test-time augmentation is applied by generating multiple shifted views of the same input:

$$\hat{y}^{(i)} = f(x^{(i)}), \tag{C.19}$$

Table C.2: AST-based downstream classification setup.

| **Component** | **Setting** |
|---|---|
| *Model Architecture* | |
| Backbone | AST (Audio Spectrogram Transformer) |
| Feature layer | Block 7 (CLS token) |
| Feature dimension | 768 |
| Classifier | Linear ($768 \rightarrow 1$) |
| Dropout | 0.65 |
| *Input Processing* | |
| Sampling rate | 16 kHz |
| Audio duration | 6 seconds |
| Channels | Mono |
| High-pass filter | 50 Hz |
| Silence trimming | Energy-based (top_db = 30) |
| Representation | AST feature extractor (log-Mel based) |
| *Training Setup* | |
| Batch size | 32 |
| Optimizer | AdamW (SAM for training phase) |
| Learning rate (backbone) | $2 \times 10^{-6}$ |
| Learning rate (head) | $2 \times 10^{-5}$ |
| Weight decay | 0 |
| Scheduler | Cosine annealing |
| Loss | Binary cross-entropy (logits) |
| *Augmentation* | |
| Feature augmentation | Time masking, frequency masking |
| Mix up | Enabled (probabilistic) |
| *Evaluation* | |
| Test-time augmentation | 3 views (original, ±500 ms shift) |
| Prediction aggregation | Mean over views |
| Threshold selection | Validation-based (Youden's J) |
| Metrics | F1, AUC, Sensitivity, Specificity |

and aggregating predictions:

$$\hat{y}_{\text{final}} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}^{(i)}. \tag{C.20}$$

The decision threshold is selected on the validation set:

$$J(\tau) = \text{TPR}(\tau) - \text{FPR}(\tau), \tag{C.21}$$

$$\tau^* = \arg\max_{\tau} J(\tau), \tag{C.22}$$

and applied unchanged during test evaluation.

## D Text Prompts

Prompts are constructed from metadata fields available in the respiratory datasets, including age, sex, smoking status, and respiratory conditions such as asthma and COPD. These attributes are converted into simple text descriptions that specify the target sound and its context. The prompts provide the main conditioning signal during generation, while metadata embeddings and the memory module are used as additional inputs within the model.

### D.1 Main Evaluation Prompts

A sustained /e/ vowel sound from a 57-year-old female smoker with possible asthma and no COPD, with reduced breath support.

A dry cough from a 65-year-old male smoker with no asthma and no COPD.

A cough from a 50-year-old female non-smoker with possible asthma and no COPD.

A cough from a 74-year-old female smoker with no asthma and no COPD.

A cough from an 18-year-old female non-smoker with possible asthma and no COPD.

A sustained /a/ vowel sound from a 20-year-old male non-smoker with no asthma and no COPD.

A sustained /o/ vowel sound from a 60-year-old male smoker with no asthma and no COPD.

A cough from a 77-year-old male smoker with no asthma and no COPD.

A cough from a 60-year-old male smoker with possible asthma and possible COPD, with slightly impaired airflow.

A cough from a 73-year-old male smoker with no asthma and no COPD, with reduced breath strength.

A cough from a 58-year-old male smoker with asthma and COPD, with heavy chest involvement.

A cough from a 48-year-old male smoker with no asthma and no COPD.

A cough from a 65-year-old male smoker with COPD, with obstructed airflow characteristics.

A sustained /a/ vowel sound from a 24-year-old male non-smoker with no asthma and no COPD.

A cough from a 30-year-old female smoker with asthma and no COPD.

A cough from an 80-year-old male smoker with COPD, with weak respiratory effort.

Shallow breathing from a 54-year-old female smoker with no asthma and no COPD.

Deep breathing from a 37-year-old female smoker with no asthma and no COPD.

A cough from an 18-year-old female non-smoker with no asthma and no COPD.

A cough from a 65-year-old male smoker with COPD, with reduced airflow.

### D.2 Example Prompt Variations

A dry cough from a female smoker aged 30–40 years. A wet cough from a male non-smoker aged 40–50 years. Shallow coughing from a person with asthma. A heavy cough from an elderly person.

## E Survey Design

Participants listened to each audio sample using an embedded audio player and rated it using a questionnaire. Each sample was evaluated independently on a 5-point Likert scale. The interface displayed the waveform alongside the audio. General listeners rated overall quality and alignment with the prompt, while clinical experts additionally rated clinical usefulness. The scales were designed to separate perceptual quality from clinical relevance.

### E.1 General Listener Evaluation (Artificial Example)

**Audio Sample:**
"A sustained /a/ vowel sound produced by a 24-year-old male non-smoker with no asthma and no COPD"

**Overall Quality (OVL)**
(Consider clarity, naturalness, and overall sound quality.)

| Very Poor | Poor | Fair | Good | Excellent |
|---|---|---|---|---|

**Relevance to Text Input (REL)**
How well does this audio match the description?

| Not Relevant at All | Slightly Relevant | Moderately Relevant | Very Relevant | Highly Relevant |
|---|---|---|---|---|

### E.2 Expert Clinical Evaluation (Artificial Example)

**Audio Sample:**
"A cough from a 60-year-old male smoker with asthma and COPD, with impaired airflow"

**Overall Quality (OVL)**
(Consider clarity, naturalness, and overall sound quality.)

| Very Poor | Poor | Fair | Good | Excellent |
|---|---|---|---|---|

**Relevance to Text Input (REL)**
How well does this audio match the description?

| Not Relevant at All | Slightly Relevant | Moderately Relevant | Very Relevant | Highly Relevant |
|---|---|---|---|---|

**Clinical Relevance for Assessment (CRA)**
(Consider whether the audio contains meaningful respiratory patterns such as realistic cough structure, airflow limitation, and clinically interpretable acoustic features.)

| Not Clinically Useful at All | Slightly Useful | Moderately Useful | Very Useful | Highly Clinically Useful |
|---|---|---|---|---|