ICLR
International Conference On Learning Representations

Resp-Agent

An Agent-Based System for Multimodal Respiratory Sound Generation and Diagnosis

Closing the gap between Analysis and Generation in Clinical AI.

Pengfei Zhang · Tianxin Xie · Minghao Yang · Li Liu

HKUST (Guangzhou)

Background

Clinical Importance & Representational Challenges

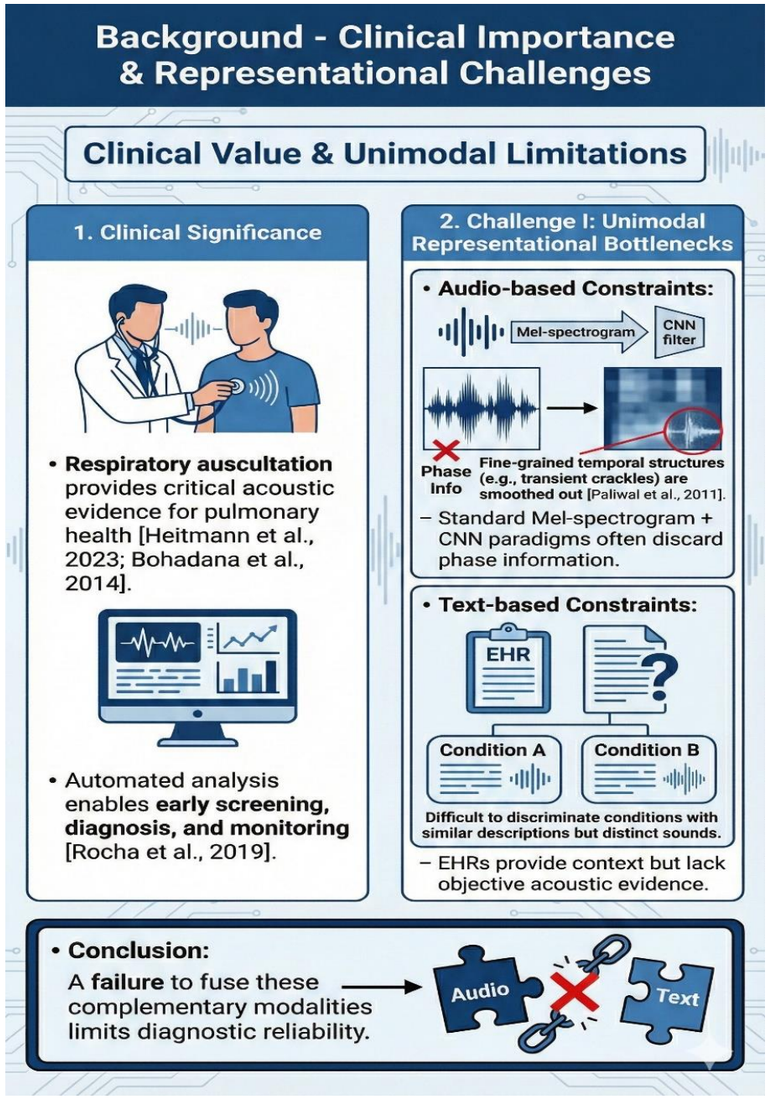
1. Clinical Significance

- Respiratory auscultation provides critical acoustic evidence for pulmonary health [1].
- Automated analysis enables early screening, diagnosis, and monitoring [2].

2. Challenge I: Unimodal Representational Bottlenecks

- Audio-based Constraints:
  - Standard Mel-spectrogram + CNN paradigms often discard phase information.
  - Fine-grained temporal structures (e.g., transient crackles) are smoothed out [3].

- Text-based Constraints:
  - EHRs provide context but lack objective acoustic evidence.
  - Difficult to discriminate conditions with similar descriptions but distinct sounds.



Consequently, the inability to leverage cross-modal synergy creates a performance ceiling.

[1] Heitmann, Julien, et al. "DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries." NPJ digital medicine 6.1 (2023): 104.

[2] Rocha, Bruno M., et al. "An open access database for the evaluation of respiratory sound classification algorithms." Physiological measurement 40.3 (2019): 035001.

[3] Paliwal, Kuldip, Kamil Wójcicki, and Benjamin Shannon. "The importance of phase in speech enhancement." speech communication 53.4 (2011): 465-494.

Background

3. Challenge II: Scarcity of Multimodal Datasets

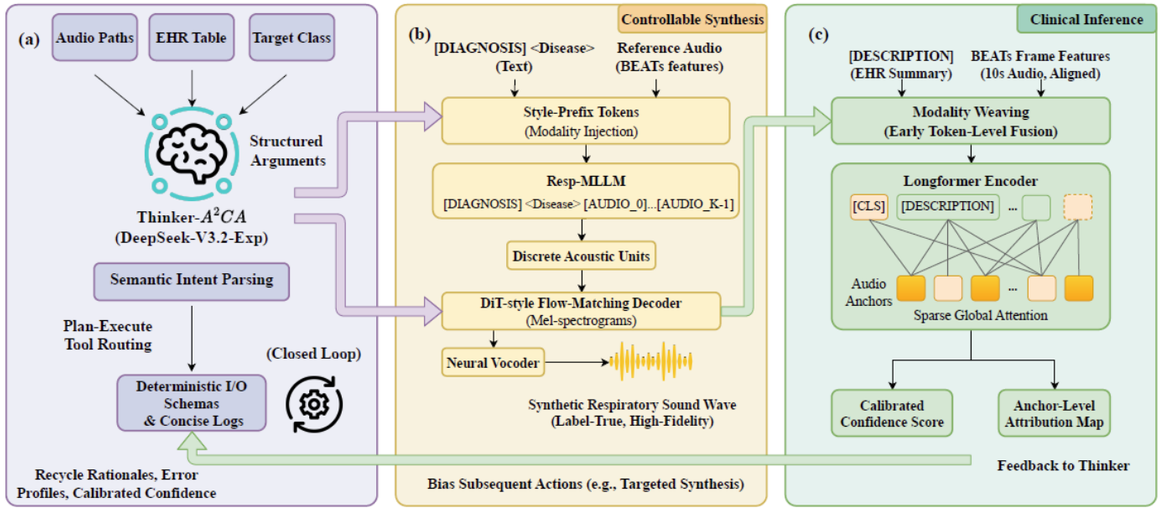
- Current Limitations: Small scale, limited disease coverage, and minimal diagnostic labels.
- The Missing Link: Lack of structured clinical text (demographics, medical history) temporally aligned with audio.
- Impact: This Impedes the training of robust multimodal large models.

The field is defined by a "bipolar" data ecosystem, necessitating domain adaptation techniques.

Category	Datasets	Characteristics	Limitations
Clinical-Grade	ICBHI, SPRSound, KAUH	Controlled environment, electronic stethoscopes, expert annotations.	Small scale (hundreds of subjects), severe class imbalance.
Crowdsourced	COUGHVID, UK COVID-19	Massive scale (25k+ recordings), diverse devices (smartphones).	High noise, weak labels, device variability artifacts.

Insight: *Pretraining on massive noisy data (Contrastive Learning) followed by fine-tuning on clinical data is the dominant paradigm.*

Overview



- **Resp-229k Benchmark:**

Establishes a new standard with a large-scale, clinically contextualized dataset comprising **229k Audio–EHR samples** (≈408 hours) across 16 diagnostic categories, featuring source-disjoint splits for rigorous OOD evaluation.

- **Controllable Resp-MLLM Generator:**

Introduces the first multimodal LLM for respiratory synthesis, utilizing **conditional flow matching** and **BEATs-derived style tokens** to disentangle disease label from acoustic style, enabling high-fidelity generation of rare disease samples.

- **Modality-Weaving Diagnoser:**

Proposes a fusion architecture that interleaves EHR text with audio via **Strategic Global Attention**, allowing the model to capture millisecond-level transient events (≈80ms resolution) while resolving long-range clinical dependencies.

- **Closed-Loop Agentic Framework:**

Bridges the data gap via the **Thinker-A²CA agent**, which actively identifies diagnostic weaknesses and schedules targeted synthesis in a curriculum, shifting from passive analysis to active **generation↔diagnosis co-design**.

Method: RespMLLM (Generator)

- **Pioneering Controllable Synthesis**

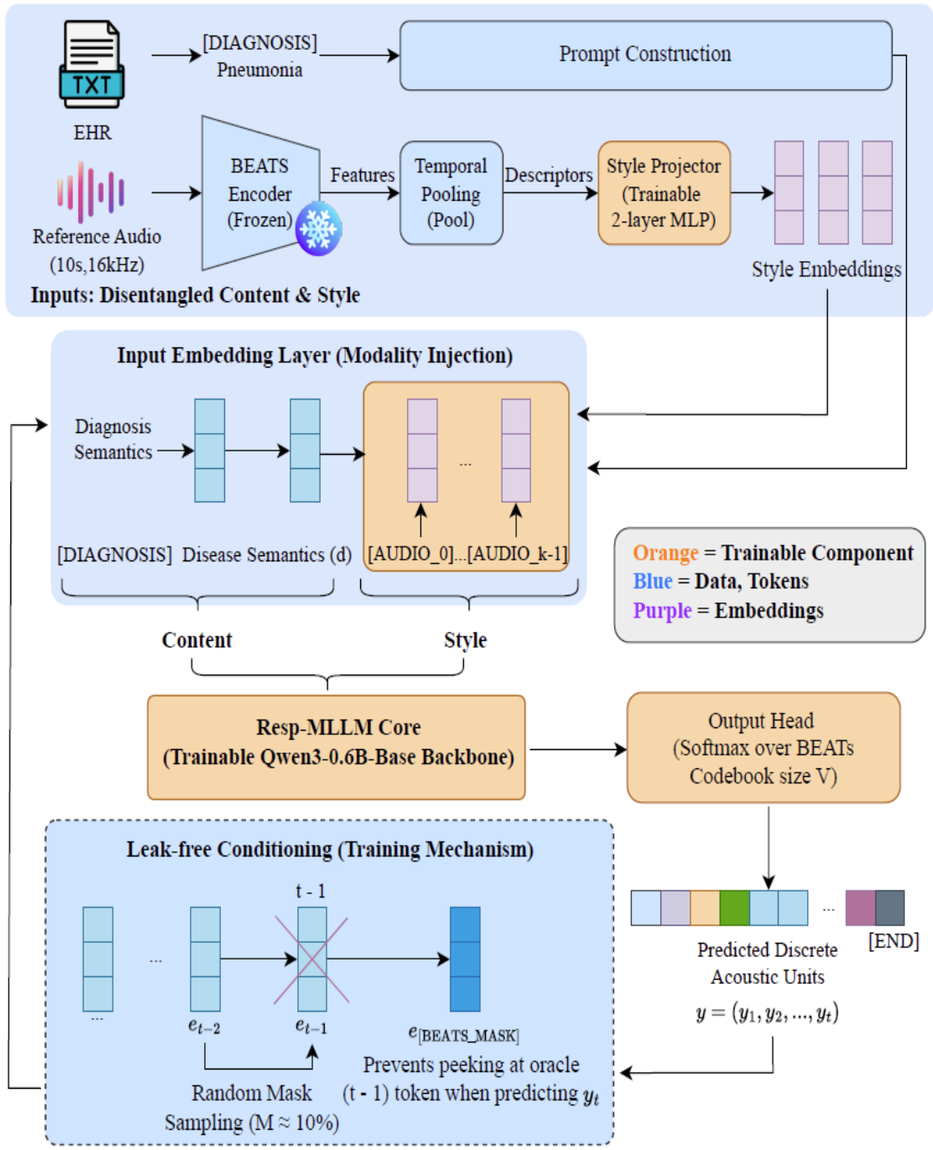
The first Multimodal LLM designed to bridge the “**data generation gap**”, replacing generic data augmentation with pathology-controlled synthesis that can generate specific disease features from clinical text descriptions.

- **Content-Style Disentanglement**

Innovatively separates "what to generate" (pathological content via text) from "how it sounds" (acoustic style via BEATs tokens), employing modality injection to strictly bind disease semantics to a reference timbre.

- **Discrete Planning with Flow Matching**

Adopts a robust two-stage architecture: the LLM autoregressively plans discrete acoustic units, while a Conditional Flow Matching (CFM) decoder reconstructs high-fidelity waveforms, ensuring phase-accurate capture of transient events.



Method: Audio Anchor (Diagnoser)

Keypoint 1: The Challenge (Addressing the Representation Gap)

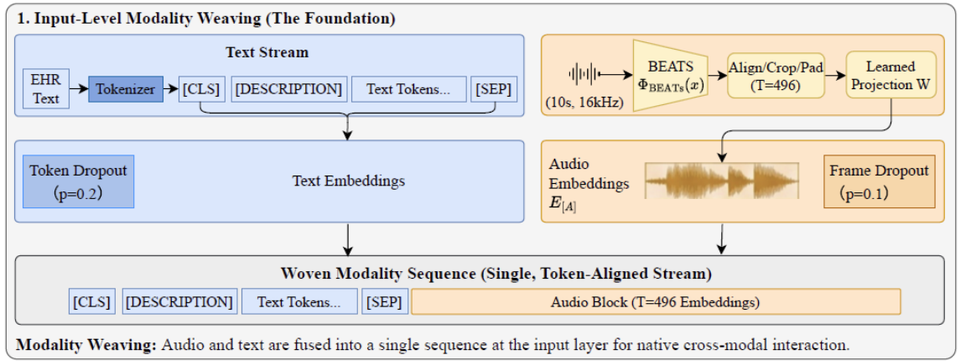
Solving the "Representation Gap": Standard audio encoders (CNNs, ASTs) often smooth out brief, low-energy events like crackles or wheezes.
The Bottleneck: Existing multimodal methods rely on "late fusion," which fails to capture fine-grained correlations between clinical text and millisecond-level acoustic transients.

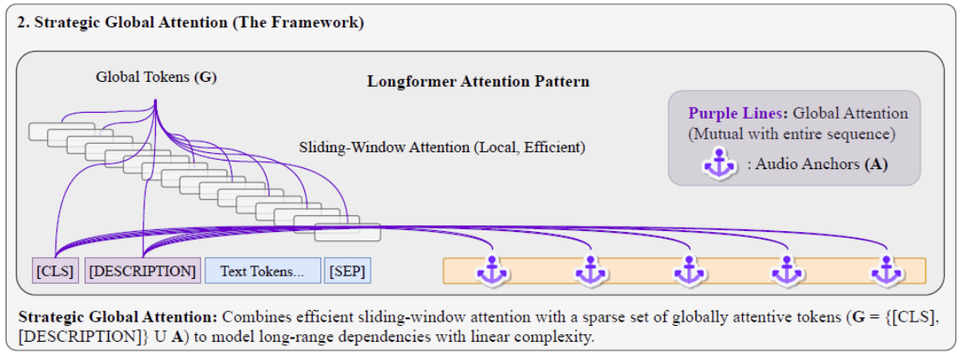
Keypoint 2: The Solution (Mechanism)

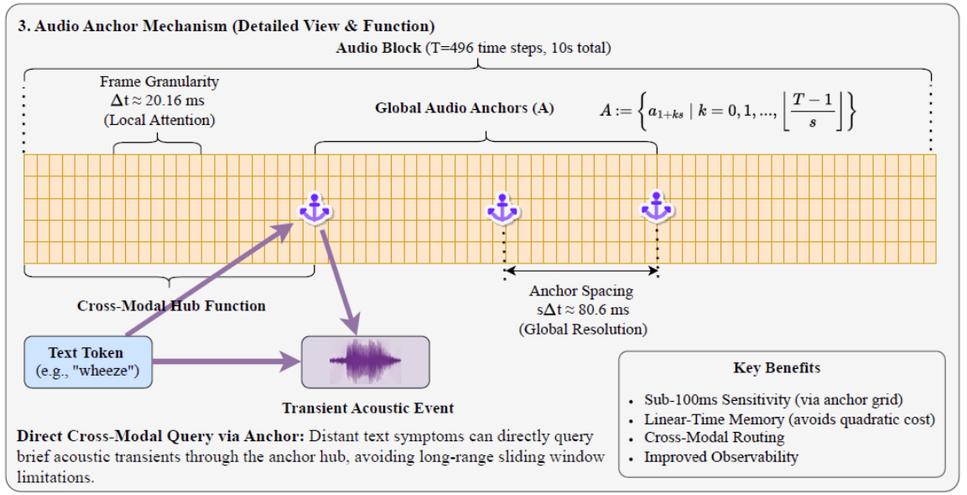
Modality Weaving with Sparse Audio Anchors: We interleave EHR text tokens directly with audio tokens at the input level, rather than processing them separately.
Strategic Global Attention: Instead of computationally expensive full attention, we place "Audio Anchors" every ~80ms. These act as global hubs that allow clinical text to efficiently query specific acoustic moments.

Keypoint 3: The Advantage (Result)

High-Resolution Grounding at Linear Complexity: This design achieves sub-100ms temporal sensitivity, enabling the model to "pin" textual symptoms (e.g., "dry cough") to fleeting respiratory events.
Efficiency: It enables long-context cross-modal reasoning without the quadratic computational cost of standard Transformers.







Rsp-229K Benchmark

(a) Resp-229k split statistics (effective samples)

Split	#Files	Hours	Mean (s)	Max (s)
Train	196,654	341	6.2	86
Valid	16,931	31	6.6	71
Test	15,516	36	8.4	30
Total	229,101	408	6.4	86

(b) Source datasets for the curation of RESP-229K

Name	Role	Device	Sample Rate (kHz)	Mean Duration (s)
UK COVID-19	Train/Validation	Microphone	48	5.9
ICBHI	Train/Validation	Stethoscope	4–44.1	22.2
SPRSound	Train/Validation	Stethoscope	8	11.0
COUGHVID	Test	Microphone	48	6.9
KAUH	Test	Stethoscope	4	15.0

- **Resp-229k Benchmark:** A unified corpus of **229,101 quality-controlled samples** across **16 classes**, featuring a rigorous out-of-domain evaluation protocol to test true generalization.

- **Standardized Clinical Summaries:** Introduces context-rich textual supervision by synthesizing heterogeneous metadata (e.g., demographics, symptoms, acoustic events) into cohesive, adaptive paragraphs.

- **Automated Data-to-Text Pipeline:** Leverages **DeepSeek-R1-Distill-Qwen-7B** to programmatically convert diverse source fields into consistent, schema-grounded annotations without high-cost manual effort.

Experimental Result

Keypoint 1: Establishing a New State-of-the-Art on ICBHI

"Resp-Agent outperforms all prior baselines by a significant margin, achieving a new SOTA Score of 72.7%. "

- **Evidence:** On the standardized ICBHI benchmark, Resp-Agent surpasses the previous best method (Dong et al., 2025) by **+5.15%** in overall Score.
- **Takeaway:** This demonstrates that our LLM-driven, agentic approach is superior to traditional CNN and Transformer-based audio classifiers.

Keypoint 2: Solving the 'Sensitivity Gap' in Disease Diagnosis

"We achieve a breakthrough in Sensitivity (66.1%), drastically reducing missed diagnoses compared to existing methods. "

- **Evidence:** While competitors struggle with sensitivity (typically 30-50%), Resp-Agent achieves **66.10%**, balancing high specificity with the ability to detect actual pathological events.

- **Takeaway:** This high sensitivity confirms that our model effectively captures transient acoustic events (like crackles/wheezes) rather than biasing toward the majority "healthy" class.

Table 2: The RSC performance on the ICBHI dataset with the official 60–40% train–test split. Here, in the Pretraining Data column, IN, AS, LA, HF and SPR refer to ImageNet(Deng et al., 2009), AudioSet(Gemmeke et al., 2017), and LAION-Audio-630K(Wu et al., 2023), HF_Lung_V1(Hsu et al., 2021) and SPRSound respectively. * denotes the previous state-of-the-art ICBHI Score. The **SOTA** and second best results are highlighted by the bold characters and underlines.

Method	Backbone	Pretraining Data	S_p (%)	S_e (%)	Score (%)
SE+SA(Yang et al., 2020)	ResNet18	-	81.25	17.84	49.55
LungRN+NL(Ma et al., 2020)	ResNet-NL	-	63.20	41.32	52.26
RespireNet(Gairola et al., 2021)	ResNet34	IN	72.30	40.10	56.20
Chang *et al.*(Chang et al., 2022)	CNN8-dilated	-	69.92	35.85	52.89
Ren *et al.*(Ren et al., 2022)	CNN8-Pt	-	72.96	27.78	50.37
Wang *et al.*(Wang & Wang, 2022)	ResNeSt	IN	70.40	40.20	55.30
Late-Fusion(Pham et al., 2022)	Inc-03 + VGG14	IN	<u>85.60</u>	30.00	57.30
Nguyen *et al.*(Nguyen & Pernkopf, 2022)	ResNet50	IN	79.34	37.24	58.29
Moummad *et al.*(Moummad & Farrugia, 2023)	CNN6	AS	75.95	39.15	57.55
Bae *et al.*(Bae et al., 2023a)	AST	IN+AS	81.66	43.07	62.37
Kim *et al.*(Kim et al., 2023)	AST	IN+AS	80.72	42.86	61.79
Kim *et al.*(Kim et al., 2024a)	AST	IN+AS	79.87	43.55	61.71
Kim *et al.*(Kim et al., 2024b)	AST	IN+AS	82.47	40.55	61.51
BTS (Kim et al., 2024c)	CLAP	LA	81.40	45.67	63.54
Wang *et al.*(Wang et al., 2024)	HTS-AT	IN+AS	79.61	48.77	64.19
MVST (He et al., 2024b)	AST	IN+AS	81.99	<u>51.10</u>	66.55
Dong *et al.*(Dong et al., 2025)	AST	IN+AS	**85.99**	49.11	67.55*
Resp-Agent[Ours]	LLM+Longformer	HF+SPR	79.29	**66.10**	**72.70**

Table 6: Performance comparison of text-only, audio-only, and multimodal models on the cross-domain test set (original, imbalanced data). Text-only Transformers improve over LSTM but fail to match audio-only baselines, justifying the need for multimodal fusion.

Model	Modality	Accuracy	Macro-F1
LSTM (main paper baseline)	Text	0.0912	0.0401
BERT-base	Text	0.1420	0.0710
RoBERTa-base	Text	0.1513	0.0742
Longformer-base (text-only)	Text	0.1585	0.0813
Conformer (audio-only)	Audio	0.7200	0.1935
Resp-Agent Diagnoser (Ours)	**Audio + Text**	**0.8494**	**0.2118**

Experimental Result

Table 7: Ablation on fusion strategies and audio encoder backbones (original, imbalanced data). Deep Modality Weaving outperforms shallow fusion methods, and the choice of fusion architecture outweighs marginal gains from changing the audio backbone.

Model / Fusion Strategy	Modality	Accuracy	Macro-F1
Conformer (audio-only, main paper)	Audio	0.7200	0.1935
Whisper-Small (audio-only)	Audio	0.7310	0.2010
Conformer + LSTM (Concat-MLP)	Audio + Text (Late)	0.8012	0.2003
Conformer + BERT (Concat-MLP)	Audio + Text (Late)	0.8124	0.2040
Conformer + BERT (Logit-Voting)	Audio + Text (Late)	0.8043	0.1992
Resp-Agent Diagnoser (Ours)	**Audio + Text (Weaving)**	**0.8494**	**0.2118**

Table 4: Compact summary of Generator content–style disentanglement

Generator disentanglement (Exp. 6)			
Test	Style-Sim ↑	P-Acc ↑	FAD ↓
Style-swap (avg over 4 styles)	0.91	97.9%	1.18
Content-swap (avg over 4 labels)	0.93	96.1%	1.19
Diagnoser ablations on Test-CD (Exp. 7)			
Config	Acc	Macro-F1	
Late Fusion, Raw Metadata, no anchors	0.780	0.145	
Late Fusion, LLM EHR, no anchors	0.790	0.160	
Modality Weaving, Raw Metadata, no anchors	0.640	0.175	
Modality Weaving, LLM EHR, no anchors	0.650	0.189	
Modality Weaving, Raw Metadata, anchors	0.835	0.195	
Full Resp-Agent Diagnoser (LLM EHR + anchors)	**0.849**	**0.212**	

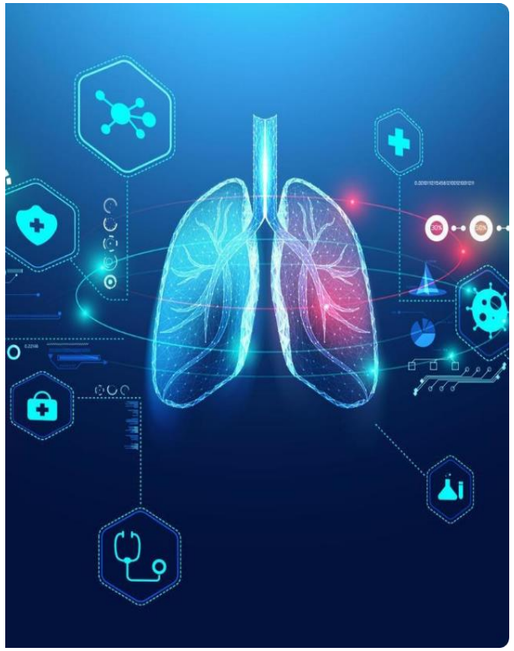
Keypoint 3: 'Modality Weaving' Proves Superior to Standard Fusion

"Our deep 'Modality Weaving' architecture outperforms standard Late Fusion, proving that *how* we fuse data matters more than the backbone size. "

• **Evidence:** Ablation studies reveal that standard concatenation (Conformer+BERT) only yields 0.20 Macro-F1. In contrast, our Weaving mechanism achieves **0.2118 Macro-F1** and **0.8494 Accuracy**.

• **Takeaway:** The gain stems from our unique architecture that forces early interaction between clinical text and audio tokens, rather than just summing their outputs.

Conlusion



1. Motivation: Two Fundamental Disconnects

- **The Representation Gap:** Spectrogram compression discards transient acoustic events and clinical context.
- **The Data Gap:** Severe class imbalance and data scarcity in medical audio.

2. Core Framework: Thinker-A²CA

- **Active Adversarial Curriculum Agent:** Acts as the central system controller.
- **Closed-Loop Mechanism:** Actively identifies diagnostic weaknesses -> schedules targeted synthesis.

3. Architectural Innovations

- **Modality Weaving Diagnoser (Bridging Representation Gap):**
  - Interleaves EHR with Audio tokens beyond standard fusion.
  - Strategic Global Attention: Captures long-range dependencies while retaining millisecond-level transient details.

- **Flow Matching Generator (Bridging Data Gap):**
  - Retools text-only LLM via Modality Injection.
  - Programmatic Synthesis: Decouples pathology from style to generate high-fidelity, hard-to-diagnose samples.

4. Contributions

- **Resp-229k Benchmark**: 229k recordings paired with LLM-distilled clinical narratives.
- **Impact:** Consistent SOTA performance; robust and deployable respiratory intelligence.