BREATH: A Bio-Radar Embodied Agent for Tonal and Human-Aware Diffusion Music Generation

Yunzhe Wang、Xinyu Tang、Zhixun Huang、Xiaolong Yue、Yuxin Zeng （MiLM Plus, Xiaomi Inc.）

Abstract

We present a multimodal system for personalized music generation that integrates physiological sensing, LLM-based reasoning, and controllable audio synthesis. A millimeter-wave radar sensor non-invasively captures heart rate and respiration rate. These physiological signals, combined with environmental state, are interpreted by a reasoning agent to infer symbolic musical descriptors, such as tempo, mood intensity, and traditional Chinese pentatonic modes, which are then expressed as structured prompts to guide a diffusion-based audio model in synthesizing expressive melodies. The system emphasizes cultural grounding through tonal embeddings and enables adaptive, embodied music interaction. To evaluate the system, we adopt a research-creation methodology combining case studies, expert feedback, and targeted control experiments. Results show that physiological variations can modulate musical features in meaningful ways, and tonal conditioning enhances alignment with intended modal characteristics. Expert users reported that the system affords intuitive, culturally resonant musical responses and highlighted its potential for therapeutic and interactive applications. This work demonstrates a novel bio-musical feedback loop linking radar-based sensing, prompt reasoning, and generative audio modeling.

Agent Architecture

USER

I'm going to sleep now.
Can you play something for me?

STRUCTURED INPUT

- HeartRate: 87 bpm
- Respiration: 17 bpm
- Time: 23:10
- Temperature: 28 °C
- Status: Preparing for sleep

Context-Aware Music Composition Reasoning

<Think>

Step 1: Interpret physiological signals

Heart rate = 87 bpm → Mildly elevated, suggesting the user is still somewhat alert or experiencing pre-sleep arousal.
Respiration rate = 17 bpm → Normal and steady, indicating no immediate stress.

Step 2: Consider contextual cues

Time = 23:10 → Late evening, typical sleep preparation time.
Temperature = 28°C → Warm but not extreme; may encourage a preference for soft, enveloping sonic textures.
User status = Preparing for sleep → Requires music that guides the listener into a restful state, balancing engagement and calming effect.

Step 3: Map user state to musical decisions

Use a **slow tempo (~65 BPM)**: slightly more active than sleep-state music, but still slow enough to encourage down regulation.
Choose **a calm and immersive genre**, such as **ambient**.
Instrumentation should remain gentle but slightly richer than in a deep-sleep state - e.g., **warm pads, soft mallets**, or **plucked textures** that fade smoothly.
Select the **Yu (羽)** mode to foster for transition.

I will output the musical plan in a structured format with the following four components:

</Think>

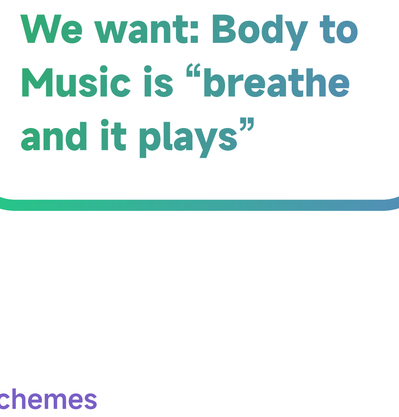
Structured Output

<Response>

Tempo: 65 BPM
Style/Genre: Ambient
Instruments: Warm ambient pads, soft mallets, gentle plucked strings
Tonal Mode: Yu

</Response>

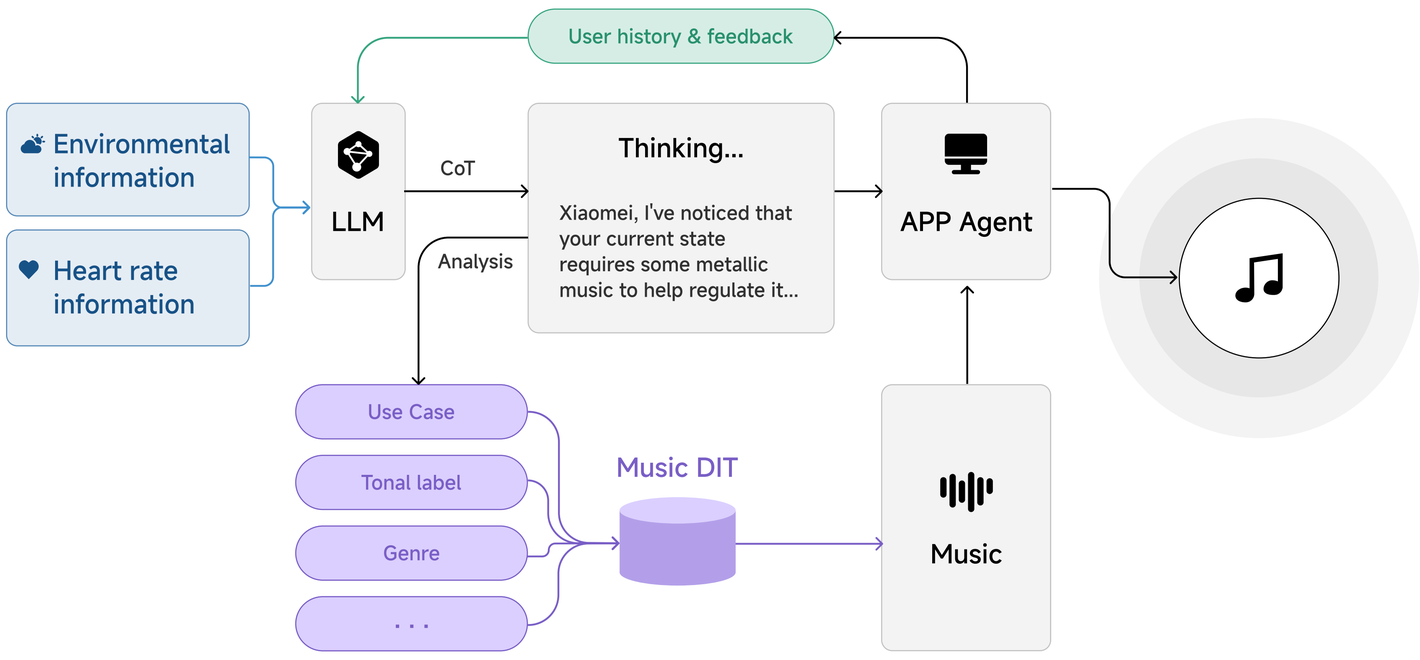
Why Hardware Matters ?



Text to Music is “ask and receive”

We want: Body to Music is “breathe and it plays”

System Overview



Experiments

Tonal accuracy under three conditioning schemes

Experiment	Accuracy
With Tonal Embedding and Prompt With Prompt	87%
With Prompt Only	22%
With Tonal Label in Prompt	43%

The result verifies that concatenating a learnable pentatonic vector at the input noise stage effectively biases the entire generation trajectory toward the desired pentatonic mode.

Generation Model

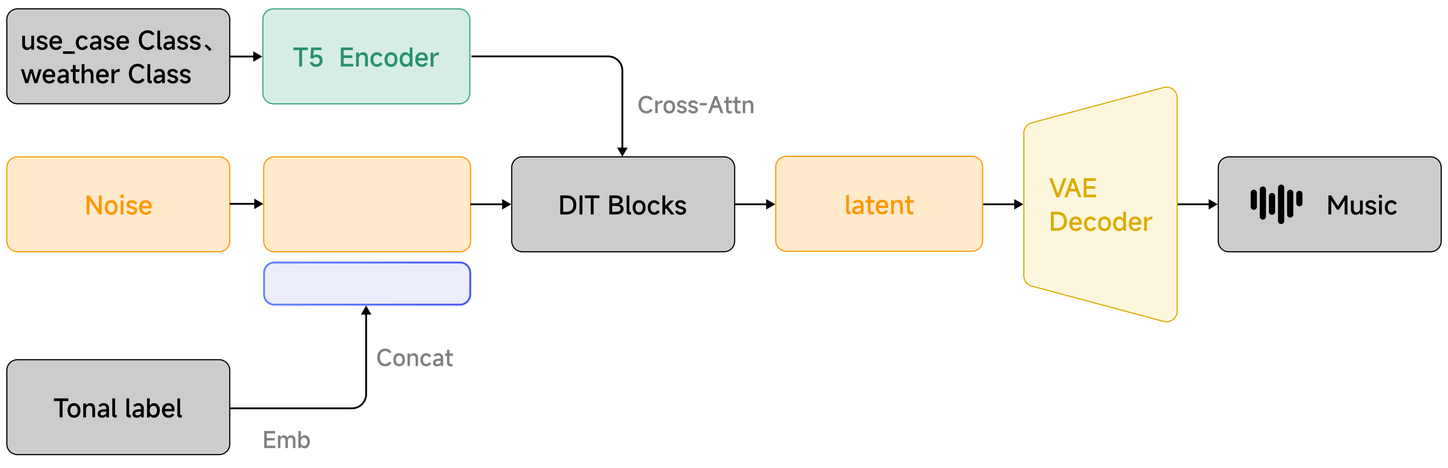
We don’t want “any music”,
We want “music that feels Chinese”

Datasets

- 500K instrumental music tracks from diverse genres
- 50K instrumental Chinese pentatonic music excerpts
- Extract textual tags, such as instrument, style, tempo with multimodal LM
- Extract tonal condition with a CNPM classifier

DIT Model

The pentatonic label is squeezed into a 1×d learnable vector and simply concatenated to the input noise, locking the melody into the pentatonic pitch space from the very first denoising step.



Future

- Deploy dense mmWave radar arrays for sub-10 cm localization and synchronized multi-person HR/RR capture, unlocking spatially-aware, group-adaptive music control.
- Curate a multimodal dataset (radar ↔ facial ↔ audio) with frame-level alignment to fuel conditional generation research.
- Streamline DiT inference to ⩽ 300 ms end-to-end, delivering truly real-time, body-driven musical feedback.

More demo in Github :

https://youknowwyz.github.io/BREATH_Music/