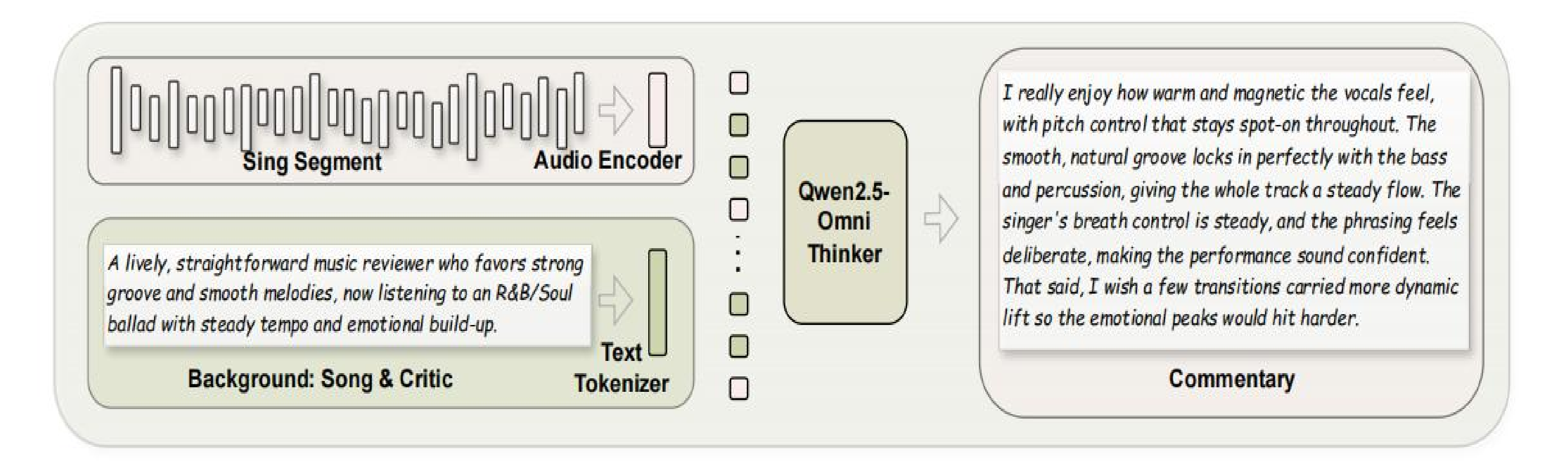
Generative Feedback for Singing Voice Synthesis Evaluation

Introduction

Singing Voice Synthesis (SVS) has advanced rapidly, yet evaluation remains limited by scalar reward models that lack interpretability and overlook expressive dimensions.

We propose a **generative feedback framework** that produces **natural language commentary**, enabling **interpretable, multi-dimensional evaluation** trained on both **synthetic MLLM** reviews and **authentic** human reactions.

Framework



A lively, straightforward music reviewer who favors strong groove and smooth melodies, now listening to an R&B/Soul ballad with steady tempo and emotional build-up.

I really enjoy how warm and magnetic the vocals feel, with pitch control that stays spot-on throughout. The smooth, natural groove locks in perfectly with the bass and percussion, giving the whole track a steady flow. The singer's breath control is steady, and the phrasing feels deliberate, making the performance sound confident. That said, I wish a few transitions carried more dynamic lift so the emotional peaks would hit harder.

Input

- Singing audio segments
- Textual metadata: song attributes + critic persona profiles

Model

- Built on Qwen2.5-Omni-7B thinker module
- Fine-tuned with LoRA for efficiency and generalization

Output

- Multi-dimensional feedback covering melody, rhythm, creativity, expressiveness, overall impression
- Commentary shaped by musical content and critic persona

Inference

- Commentary generated using top-p sampling to balance coherence and diversity

Evaluation Protocol

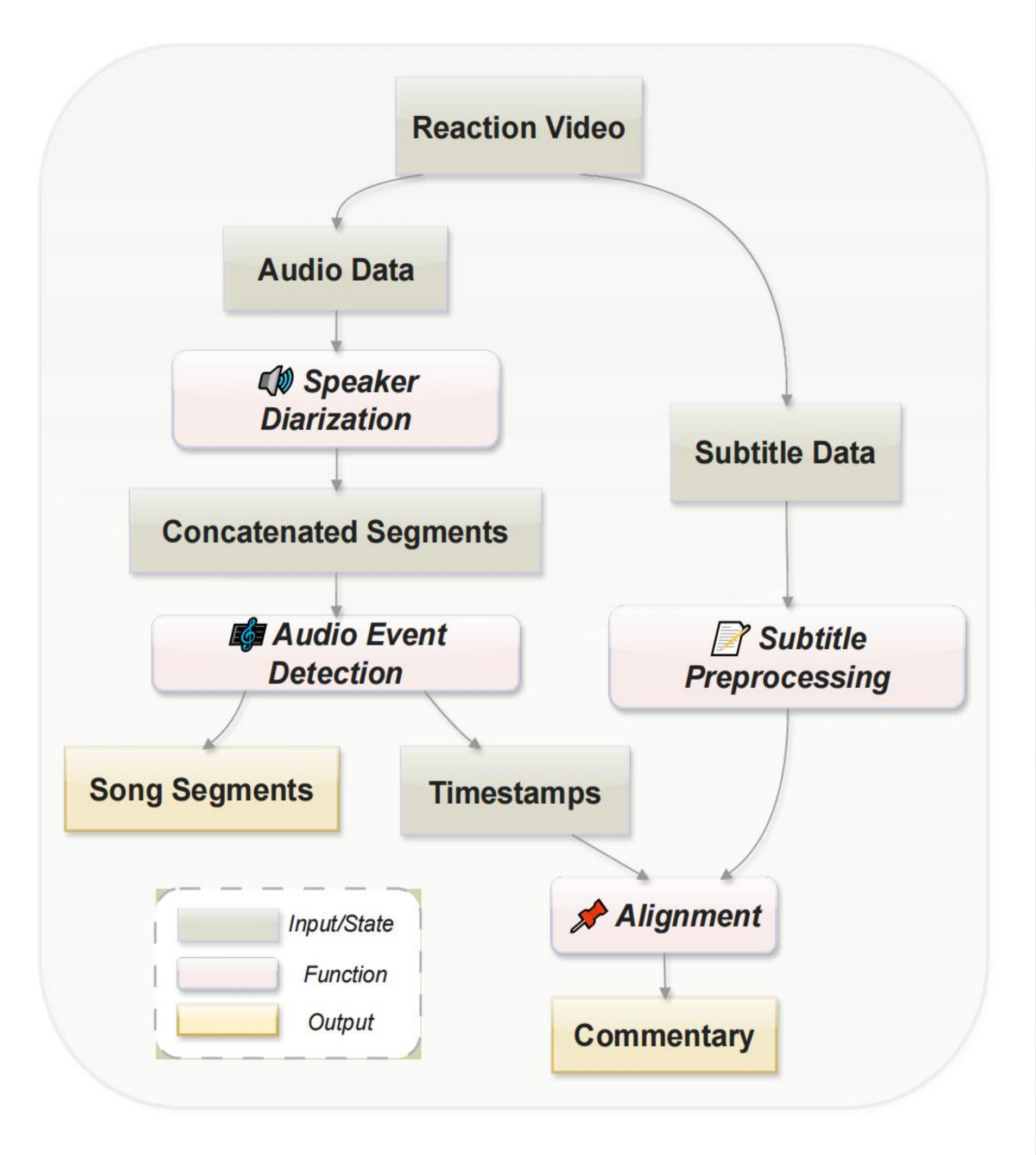
Evaluating commentary with an LLM-based benchmark:

- **Musical QA** for knowledge
- **Completeness** of coverage
- **Precision** against metadata
- **Novelty** of insights

This provides multi-dimensional, interpretable evaluation.

Dataset Construction

We construct a multimodal dataset where each sample pairs a **10–60s singing segment** with **contextual metadata**, including song attributes and critic personas.



MLLM-generated reviews: Synthetic commentary across diverse genres, guided by critic personas → systematic coverage of vocal quality.

Human reaction data: From YouTube reaction videos → authentic judgments and diverse real-world styles.

Experiments & Results

Model Variant	Validation Dataset Loss		LLM-based Reaction Benchmark			
	MLLM ↓	Reaction ↓	QA ↑	Completeness ↑	Precision ↑	Novelty ↑
Gemini-2.5-Flash [23]	-	-	52.8%	0.606	0.917	0.523
Qwen2.5-Omni-7B (Pretrained)	2.532	2.419	22.9%	0.832	0.604	0.688
Fine-tuned (SFT+LoRA)	1.882	1.499	65.7%	0.937	0.669	0.813

Main Results

Fine-tuned model reduces **validation loss**: 2.532 → 1.882 (MLLM), 2.419 → 1.499 (reaction).

QA accuracy improves from 22.9% → 65.7%.

Completeness rises to 0.937, with clear gains in Novelty and stronger Precision.

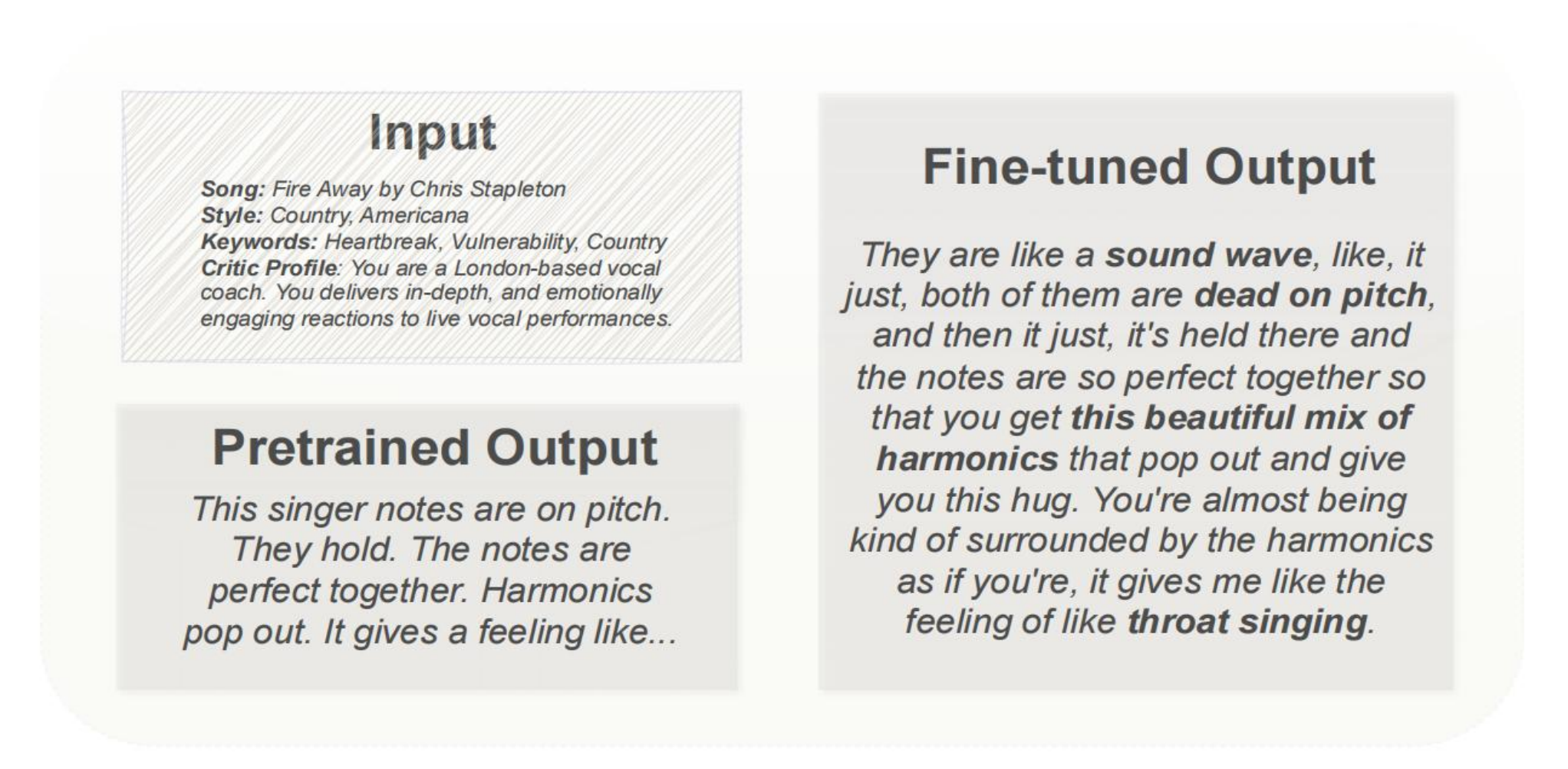
Outperforms Gemini-2.5-Flash in multiple dimensions.

Ablation Study

Using **only** synthetic data → better coverage but weak realism.

Using **only** reaction data → authentic but less systematic.

Combining both subsets yields the best overall performance, confirming their complementarity.



Input

Song: *Fire Away by Chris Stapleton*
Style: *Country, Americana*
Keywords: *Heartbreak, Vulnerability, Country*
Critic Profile*: You are a London-based vocal coach. You delivers in-depth, and emotionally engaging reactions to live vocal performances.*

Pretrained Output

This singer notes are on pitch. They hold. The notes are perfect together. Harmonics pop out. It gives a feeling like...

Fine-tuned Output

They are like a ***sound wave****, like, it just, both of them are* ***dead on pitch****, and then it just, it's held there and the notes are so perfect together so that you get* ***this beautiful mix of harmonics*** *that pop out and give you this hug. You're almost being kind of surrounded by the harmonics as if you're, it gives me like the feeling of like* ***throat singing****.*

Model Variant	Validation Dataset Loss	
	MLLM ↓	Reaction ↓
Qwen2.5-Omni-7B	2.532	2.419
Fine-tuned (SFT+LoRA)	1.882	1.499
w. only MLLM dataset	1.809	1.832
w. only Reaction dataset	2.057	1.394
w. unfiltered data	2.262	1.951

Conclusion & Future Research

We introduce the first **generative feedback framework** for Singing Voice Synthesis (SVS) evaluation, producing **natural language commentary** instead of scalar scores. This approach enables **interpretable, multi-dimensional assessment** and leverages both **synthetic MLLM reviews** and **authentic human** reactions for robustness.

Experiments show clear gains in **accuracy, completeness, and novelty**, surpassing baselines. Our framework not only enhances SVS evaluation but also opens paths toward **interactive control** and **RLHF-driven optimization** in broader music generation tasks.

Looking ahead, we aim to extend this framework to broader music generation tasks, enable interactive user control, and integrate it with RLHF pipelines for real-time optimization.