# GENERATIVE FEEDBACK FOR SINGING VOICE SYNTHESIS EVALUATION


## ABSTRACT

Singing voice synthesis (SVS) has advanced significantly, enabling models to generate vocals with accurate pitch, and consistent style. As these generative capabilities improve, the need for reliable evaluation and optimization becomes increasingly critical. However, current methods like reward systems often rely on single numerical scores, struggle to capture complex dimensions such as phrasing or expressiveness, and require costly annotations, limiting interpretability and generalization. To address these issues, we introduce a generative feedback (i.e., reward model) framework that outputs natural language commentaries rather than a scalar value, providing interpretable and multi-dimensional evaluation signals for SVS. Our approach traines a reward model capable of generating text commentary across melody, rhythm, creativity, and overall quality, integrating audio with contextual metadata within a pretrained model to yield multi-dimensional and interpretable feedback. Training is conducted on a complementary dataset that combines commentary generated by MLLMs with authentic human feedback from real-world reactions, capturing both large-scale diversity and real-world evaluation patterns. Experiments demonstrate that this framework not only improves the style consistency, and expressiveness of SVS evaluation, but also delivers stronger interpretability and better generalization and diversity compared to conventional baselines.


## 1. INTRODUCTION

Recent advancements in singing voice synthesis have experienced rapid development [1–5]: current systems are increasingly capable of producing vocal performances with increasingly accurate pitch, precise rhythm and stylistic consistency based on given inputs. Effectively evaluating generated results and leveraging these evaluations to guide subsequent model optimization remains a key challenge in this field [6]. Feedback (e.g., predefined criterion rules [7–9] and learned reward models [6, 10]) play a crucial role in this process, as they provide essential signals during evaluation and training, enabling control over generation quality and facilitating iterative improvement. A prominent category of reward designs for singing generation is based on music theory and rule systems [7, 11, 12], which define concrete functions in terms of rhythm, tonality, and other interpretable aspects. While conceptually straightforward, the generalizability, as the ability to maintain performance for unseen singers or novel musical styles, is often limited. Consequently, existing SVS system [1, 4] struggle to capture high-level characteristics and nuanced artistic expressiveness which are crucial for achieving authentic and compelling vocal synthesis.

To address this challenge, multifaceted handcrafted reward designs construct composite metrics on semantic dimensions—such as emotional expression [8] and style alignment [9]. In addition, neural networks have been employed to extract these semantic features automatically for evaluation [13]. Learning rewards from human preferences [14, 15] is also a emerging approach for modeling this procedure. With these SVS-oriented reward functions, reinforcement learning algorithms such as PPO [16] have been employed to fine-tune singing voice synthesis models. This paradigm is guided by a reward model [6, 10] that provides scalar-valued feedback signals quantifying aspects such as pitch accuracy and style consistency. Subsequently, users can directly influence the reward model by selecting preferred segments, enabling explicit specification of desired model output.

However, several common issues exist across these reward systems. First, the reward output is typically a single numerical score, which fails to adequately capture the multi-dimensional nature of singing quality [17, 18]. Without a breakdown across these dimensions, the score restricts interpretability and hinders statistical analysis of representative factors, such as principal components, variance ranges, or dimension-wise contributions. This makes it more difficult to explain differences in scores and to derive actionable optimization signals for training. Second, conventional reward model necessitates explicit definitions for each dimension. Yet, certain aspects of singing, for example, phrasing flow or tempo flexibility, are inherently difficult to quantify objectively. Some approaches that leverage semantic alignment between text and music [19] provide a potential direction, but robust automatic modeling of these subtle dimensions remains challenging. Third, prevailing reward model is predominantly trained under supervised learning and therefore depends heavily on large volumes of high-quality annotated data. This requires significant resources, domain expertise, and consistent quality control. Such demand is particularly problematic in the audio domain due to its inherent complexities; for instance, inter-annotator disagreement like note boundaries can introduce label noise that misdirects model training.

**Table 1**. Comparison of the two generated dataset types.

| Data Feature | MLLM-generated Data | Human Reaction Data |
|---|---|---|
| **Audio** | High-fidelity clean song clips | In-the-wild noisy clips |
| **Text** | MLLM-generated comments | Human review transcripts |
| **Critic Style** | Prompt-controlled persona | Natural authentic expression |
| **Quality** | Systematic & completeness | Fragmented & diverse |
| **Primary Use** | Performance coverage | Authenticity & stylization |

Motivated by these considerations, we propose an interpretable, generative reward modeling framework to provide feedback for evaluating singing voice synthesis. Unlike conventional scalar score approaches, our reward model produces natural language commentary assessing outputs across multiple dimensions: content, style, structure, and other aspects. This framework not only enhances interpretability and increases evaluation coverage, but also facilitates user interaction via a language-level interface for aesthetic control. Specifically, our approach takes an singing voice audio together with an associated text as input, where the text is formed by concatenating two elements: background information about the music segment, and stylistic profile of the human/LLM critic. These inputs are processed by an audio-text understanding model, which generates a organized commentary - a multi-dimensional, text-based feedback covering aspects such as melody, creativity, and overall auditory impression. Our training data consists of two complementory sources: 1) raw singing segments associated with text generated by a multimodal large language model (MLLM) that this source enables data production with systematic commentary styles and rich textual content; and 2) singing segments extracted from reaction videos in which individuals listen to and comment on music in real time, associated with corresponding texts combined with song information and critic's profiles. Based on these datasets, we perform supervised fine-tuning (SFT) [20] on the Qwen2.5-Omni-7B [21] pretrained model. We apply low-rank adaption (LoRA) [22] to all linear layers of the thinker module for efficiency, aiming to suppress overfitting and enhance the generalization. After training, LoRA weights are merged back into the thinker module, and inference is performed autoregressively using top-p sampling to balance coherence and diversity. To evaluate our framework in realistic settings, we also design an LLM-based benchmark that integrates music-domain knowledge and audio analysis to assess singing-review quality. It scores on objective musical knowledge, completeness, factual accuracy, and novelty. Through this generative feedback, we obtain both textual commentary that can be used to train the generation model or support downstream tasks. This innovative paradigm integrates interpretability, multi-dimensional evaluation, and broad applicability into a framework for evaluating singing voice synthesis tasks.

## 2. METHOD

This section details the training framework for our proposed reward model, implemented via LoRA on the Qwen2.5-Omni-7B [21] foundation model. The model operates on synthesized/collected audio–text data, enabling it to generate natural language commentary conditioned on given inputs. Our methodology comprises three core stages: dataset construction (Section 2.1), model architecture (Section 2.2), and evaluation protocol (Section 2.3).

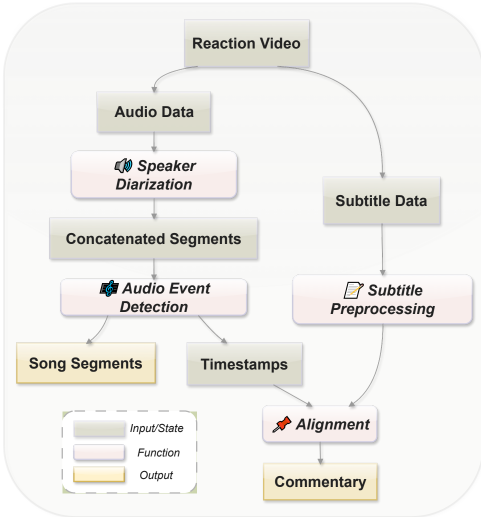


**Figure 1**. Workflow for constructing the human reaction dataset. Raw videos are split into audio and subtitle streams, followed by speaker diarization and subtitle preprocessing. Audio event detection is used to locate song segments and their timestamps. These are then aligned with the processed subtitle data to produce structured commentary outputs for downstream training and evaluation.

### 2.1 Dataset Construction

Our dataset is designed to enable the generation of high-quality singing commentary conditioned on both musical performances and contextual metadata. Adopting a unified audio-text structure, the dataset integrates two complementary sources: the first is a large set of synthetic feedback generated by MLLMs, providing systematic performance coverage; while the second is acquired from human-authored reaction videos, capturing authentic judgments to enhance realism. The differences between them are summarized in Table 1. The following sections outline the common data structure to clarify their composition.

#### *2.1.1 Data Structure*

Each dataset sample comprises a 10-60 second audio segment preprocessed via sampling rate normalization to ensure temporal consistency and mitigate source artifacts. The audio is associated with contextual text containing song attributes, critic profiles, and domain commentary grounded in contextual background. This unified multimodal organization not only ensures consistent model input formatting, but also facilitates multi-source data integration and cross-dataset evaluation. Crucially, contextual inputs guide commentary generation to capture both the sing voice content and the critic's characteristics. The former contains song attributes (creation background, composer/performer identities, thematic tags) enabling contex-

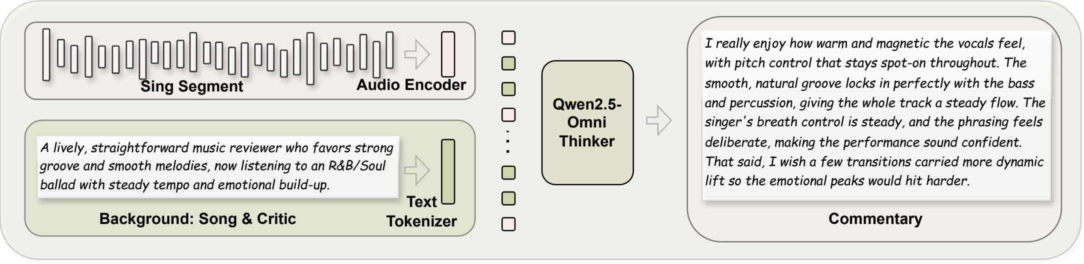


**Figure 2**. Overview of our reward model fine-tuned framework, which uses the Qwen2.5-Omni-7B thinker module's audio and text encoders with shared-attention fusion to align log-Mel spectrograms and tokenized text in a unified embedding space, enabling hierarchical cross-modal interaction and autoregressive generation of context-aware singing commentary.

tual understanding beyond acoustic signals. The latter focuses aesthetic preferences and linguistic style profiles, ensuring outputs exhibit contextual coherence, and stylistic fidelity to target personas. The following section details the acquisition methodology for each dataset type.

### *2.1.2 Category 1: MLLM-generated Data*

The first category is built from curated song segments paired with systematically generated critical commentaries. To ensure broad stylistic representation, ten diverse musical genres are included, each with representative songs. Professional critic profiles are parameterized through system prompts specifying critical tone (e.g., analytical, reactionary), linguistic patterns (encompassing rhetorical devices and phonological mimicry), genre preferences, and cultural backgrounds, enabling contextually nuanced references. Song metadata—including background, composition, and stylistic attributes—is paired with these critic profiles. This integrated input is designed to equip MLLMs with objective, expert-level music evaluation capabilities, fostering a comprehensive understanding of acoustic properties. This dataset contributes to precise textual descriptions of vocal quality assessments, thus establishing a robust foundation for subsequent training.

### *2.1.3 Category 2: Human Reaction Data*

The second dataset category derives from YouTube human reaction videos, processed into audio-text pairs via the pipeline shown in Figure 1. Raw videos undergo audio extraction using *yt-dlp* and subtitle retrieval via the *YouTube Transcript API*. Audio segments are speaker-diarized (*pyannote.audio*) and merged into utterances by speaker. An AudioSet-finetuned classifier (*MIT/ast-finetuned-audioset-10-10-0.4593*) labels utterances as singing or spoken commentary. Contiguous singing segments are then aligned with subsequent critic commentary segments. Finally, critic speech timestamps are matched to subtitles to extract review text, forming training samples including song audio with commentary. Trimmed segments are stored alongside preprocessed reviewer persona metadata (from channel introductions) and song metadata (from Wikipedia), creating multimodal samples. To ensure the data quality, we filter samples with empty subtitles/text, audio segments <10 seconds, or text <8 words, ensuring sufficient linguistic and acoustic content. This curated data enhances model capability to evaluate singing voice synthesis across diverse singing styles, speaker characteristics, and recording environments—critical for robust and diverse SVS assessment.

We investigate several alternative approaches for the reaction data pipeline but found limitations. Applying audio event detection followed by ASR is hindered by overlapping singing and commentary, leading to imprecise segmentation. Speaker-based segmentation suffered from garbled transcriptions that compromised accuracy. Aligning YouTube subtitles with speaker diarization is also unreliable due to numerous short-duration segments. Thus we adopt the previously described pipeline for this subset.

## 2.2 Model Architecture

Figure 2 illustrates the adapted training pipeline based on the Qwen2.5-Omni-7B pretrained model. While the original architecture incorporates a multimodal thinker module (integrating audio/visual encoders with a transformer-based language backbone) and a talker module for audio reconstruction, our implementation retains only audio-text capabilities since the task requires neither video input nor synthesized audio output. Audio inputs are resampled to 16kHz using *librosa* and transformed into log-Mel spectrograms, while text inputs are augmented with special tokens (*<im_start>*, *<audio>*) before tokenization. Following feature extraction, both modalities are projected into a shared latent space and processed by the language model backbone, which autoregressively generates textual output tokens. The entire network is optimized via cross-entropy loss between predicted logits and ground-truth tokens.

## 2.3 LLM-based Reaction Evaluation

To obtain a quantitative assessment in generating singing reviews, we design a comprehensive evaluation framework that leverages the average of different LLMs to score outputs across multiple dimensions. The framework consists of four complementary components. First, a multiple-choice audio QA module measures a model's musical knowledge and auditory discrimination presenting 4-8 options and calculating accuracy based on the chosen answer. Second, the completeness module employs several LLMs to score generated reviews against a set of structured criteria, ensuring coverage of all key aspects. Third, a precision

**Table 2**. Main comparison results on loss on two validation sets and the LLM-based reaction benchmark. '-' denotes unavailable metrics due to inherent constraints in calculating loss for closed-source models.

| 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 |

module checks factual consistency against verified song information, counting correct statements to compute a precision score. Finally, a novelty module rewards unique insights that go beyond common or obvious knowledge. This design with diverse evaluation tasks and LLM-based scoring allows us to capture the breadth, correctness, and originality of generated reviews, and thus serves as the evaluation framework for our reward model.

# 3. EXPERIMENTS

## 3.1 Setup

Our dataset consists of two subsets: 37 hours of MLLM-generated data and 176 hours reaction data, from which we reserve 10% of the data as a validation set. For comprehensive evaluation, we further employ the LLM-based framework introduced in Section 2.3. Within this framework, four complementary modules—QA, completeness, precision, and novelty—are defined, with each contributing an independent score. Together, these modules form a coherent assessment protocol that captures different facets of review quality. LoRA rank is configured to 8 and applied to all linear layers. Training is conducted with a per-device batch size of 2, gradient accumulation steps of 4, and the AdamW optimizer with weight decay of 0.01. The learning rate is set to `1e-4`, with a cosine learning rate schedule and a warm-up ratio of 0.1. The total number of training steps is 10000 (about 2 epochs). Each experiment is conducted on a single NVIDIA A100 GPU.

## 3.2 Main Results

Table 2 reports results on both validation losses and the LLM-based evaluation benchmark. The fine-tuned model (SFT+LoRA) substantially reduces loss compared to the pretrained base model (from 2.532 to 1.882 on the MLLM set and from 2.419 to 1.499 on the reaction set), showing stronger alignment with reference commentary. Moreover, benchmark results reveal marked improvements across multiple dimensions: QA accuracy rises from 22.9% to 65.7%, while completeness and novelty both show clear gains, reflecting reviews that are more detailed and original. As shown in Figure 3, the fine-tuned model produces coherent and contextually integrated commentary, while the pretrained output appears fragmented. Precision also improves, confirming stronger factual grounding. These benchmark gains also translate into advantages over the closed-source Gemini-2.5-Flash [23]: although Gemini achieves strong precision due to careful reasoning process, it falls behind our fine-tuned model in QA, completeness, and novelty, underscoring the strength of our framework. Together, these results demonstrate that our framework enhances the effectiveness of generated reviews.

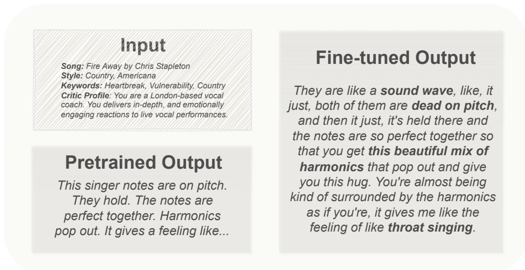


**Figure 3**. Showcase on pretrained and fine-tuned model.

## 3.3 Ablation Study

We conduct ablatios to examine the impact of different training data configurations. As shown in Table 3, using only the MLLM-generated dataset yields relatively low loss on the MLLM validation set but higher loss on the reaction set, while training with only the reaction dataset produces the opposite trend. This indicates that each data source contributes complementary strengths. Training on the unfiltered dataset leads to higher losses on both datasets, suggesting that noisy or low-information samples leads to overfit. By contrast, combining both filtered subsets in the fine-tuned model achieves the best overall results, demonstrating the importance of data quality and complementarity for improving model alignment.

**Table 3**. Ablations on different datasets for training reward models. Results are reported on two validation loss.

| 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 |

# 4. CONCLUSION

In this paper, we propose a novel framework combining natural language commentary with scalar scores to provide interpretable, multi-dimensional evaluation for SVS. Our approach trains a model to analyze melody, rhythm, and expressiveness by integrating audio and metadata, leveraging both MLLM-generated and real human feedback data.

## 5. REFERENCES


[1] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022.* AAAI Press, 2022, pp. 11 020–11 028. [Online]. Available: https://doi.org/10.1609/aaai.v36i10.21350

[2] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022.* IEEE, 2022, pp. 7237–7241. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747664

[3] Y. Yu, J. Shi, Y. Wu, Y. Tang, and S. Watanabe, "Visinger2+: End-to-end singing voice synthesis augmented by self-supervised learning representation," in *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024.* IEEE, 2024, pp. 719–726. [Online]. Available: https://doi.org/10.1109/SLT61566.2024.10832313

[4] J. Cui, Y. Gu, C. Weng, J. Zhang, L. Chen, and L. Dai, "Sifisinger: A high-fidelity end-to-end singing voice synthesizer based on source-filter model," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024.* IEEE, 2024, pp. 11 126–11 130. [Online]. Available: https://doi.org/10.1109/ICASSP48485.2024.10446786

[5] Y. Zhang, Z. Jiang, R. Li, C. Pan, J. He, R. Huang, C. Wang, and Z. Zhao, "Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds. Association for Computational Linguistics, 2024, pp. 1960–1975. [Online]. Available: https://doi.org/10.18653/v1/2024.emnlp-main.117

[6] G. Cideron, S. Girgin, M. Verzetti, D. Vincent, M. Kastelic, Z. Borsos, B. McWilliams, V. Ungureanu, O. Bachem, O. Pietquin, M. Geist, L. Hussenot, N. Zeghidour, and A. Agostinelli, "Musicrl: Aligning music generation to human preferences," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024.* OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=EruV94XRDs

[7] D. Herremans and E. Chew, "Morpheus: Generating structured music with constrained patterns and tension," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 510–523, 2019. [Online]. Available: https://doi.org/10.1109/TAFFC.2017.2737984

[8] L. Yang, S. Chou, and Y. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 324–331. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/226_Paper.pdf

[9] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, and G. Xia, "POP909: A pop-song dataset for music arrangement generation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, and T. de Reuse, Eds., 2020, pp. 38–45. [Online]. Available: http://archives.ismir.net/ismir2020/paper/000089.pdf

[10] H. Liao, H. Han, K. Yang, T. Du, R. Yang, Z. Xu, Q. Xu, J. Liu, J. Lu, and X. Li, "BATON: aligning text-to-audio model with human preference feedback," *CoRR*, vol. abs/2402.00744, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.00744

[11] N. Wang, H. Xu, F. Xu, and L. Cheng, "The algorithmic composition for music copyright protection under deep learning and blockchain," *Appl. Soft Comput.*, vol. 112, p. 107763, 2021. [Online]. Available: https://doi.org/10.1016/j.asoc.2021.107763

[12] C. Jin, F. Wu, J. Wang, Y. Liu, Z. Guan, and Z. Han, "Metamgc: a music generation framework for concerts in metaverse," *EURASIP J. Audio Speech Music. Process.*, vol. 2022, no. 1, p. 31, 2022. [Online]. Available: https://doi.org/10.1186/s13636-022-00261-8

[13] F. Carnovalini and A. Rodà, "Computational creativity and music generation systems: An introduction to the state of the art," *Frontiers Artif. Intell.*, vol. 3, p. 14, 2020. [Online]. Available: https://doi.org/10.3389/frai.2020.00014

[14] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4299–4307. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[15] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize from human feedback," *CoRR*, vol. abs/2009.01325, 2020. [Online]. Available: https://arxiv.org/abs/2009.01325

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[17] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

[18] X. Sun, Y. Gao, H. Lin, and H. Liu, "Tg-critic: A timbre-guided model for reference-independent singing evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICASSP49357.2023.10096309

[19] Y. Wang, W. Yang, Z. Dai, Y. Zhang, K. Zhao, and H. Wang, "Melotrans: A text to symbolic music generation model following human composition habit," *CoRR*, vol. abs/2410.13419, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.13419

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[21] J. H. H. H. T. H. S. B. K. C. J. W. Y. F. K. D. B. Z. X. W. Y. C. J. L. Jin Xu, Zhifang Guo, "Qwen2.5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[23] Google DeepMind, "Gemini 2.5 flash," https://deepmind.google/models/gemini/flash/, Google DeepMind, Tech. Rep., 2025, a thinking-enabled AI model capable of reasoning through intermediate steps.