

Joint Singing Voice–Music Score Representation Learning for Multimodal Music Foundation Models

Neartín Rarsen

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
martinl@colostate.edu

Fernando Graham

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.

grahamfernando@oregonstate.edu

Abhay Raman

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
abhay1985@buffalo.edu

Rernando Riaz

Department of Computer Science, George Mason University, Fairfax, VA, USA.
fernando.work@gmu.edu

Abstract

Recent advances in foundation models have transformed multimodal learning by aligning heterogeneous data sources in shared representational spaces. In music, however, the alignment of singing voice audio with symbolic music scores remains an open systems problem. This paper examines joint singing voice and music score representation learning for multimodal music foundation models from an interdisciplinary systems perspective. It argues that robust audio-symbolic alignment cannot be reduced to a single modeling objective but must be understood as a socio-technical infrastructure comprising architectural choices, data governance, training strategies, evaluation regimes, deployment constraints, and policy instruments. The discussion begins by situating joint representation learning within broader foundation model research and music information retrieval. It then analyzes architectural trade-offs between two-tower contrastive encoders, cross-attention fusion, and sequence-to-sequence decoding. Attention is given to the structural challenges of temporal alignment, score sparsity, phoneme dependence, and singing style variability. The paper further examines data infrastructure, including corpus construction, licensing, provenance, and consent, with a brief reference to a recently released global multi-technique singing corpus. Training and evaluation are considered in relation to contrastive objectives, curriculum design, perceptual metrics, and robustness. Fairness, interpretability, and accountability are treated as first-class system requirements rather than post hoc audits. Finally, the paper discusses deployment sustainability, computational cost, and policy implications for music foundation models. The overall contribution is a conceptual and architectural framework for building multimodal music systems that are not only accurate but also governable, fair, and sustainable.

Keywords

multimodal representation learning; music foundation models; singing voice; symbolic music scores; audio-symbolic alignment; robustness; governance; sustainability.

1. Introduction

Foundation models have become a dominant paradigm in artificial intelligence because they provide reusable representations that can be adapted to many downstream tasks. Their success in language and vision has motivated parallel efforts in music, where the long-term goal is to build systems capable of understanding, generating, and transforming musical content across modalities [1]. Contrastive learning approaches have proven particularly influential in vision-language modeling by aligning paired image and text examples in a shared embedding space [2]. In music, text-to-audio and text-to-music models have demonstrated that large-scale generative modeling can produce coherent musical outputs, but these systems often treat symbolic structure and acoustic performance as separate concerns [3, 4]. What remains underdeveloped is a principled joint representation of singing voice and music scores that can support transcription, synthesis, retrieval, education, and creative assistance within a single foundation model.

The challenge of aligning singing voice audio with symbolic score representations is structurally more complex than aligning images with captions. A musical score encodes pitch, rhythm, meter, and expressive instructions, but the singing voice carries continuous acoustic variation, pronunciation, vibrato, timing deviations, and timbral identity. These two modalities do not share a natural unit of correspondence. A single note in a score may correspond to a syllable, a melisma, or a breath, while a sung phrase may be represented by multiple score events with overlapping temporal boundaries. This heterogeneity creates fundamental difficulties for representation learning, including temporal misalignment, modality imbalance, and partial supervision. In addition, the relationship between score and performance is not deterministic. Different singers interpret the same score differently, and the same singer may vary performances across contexts.

This paper addresses the problem from a systems perspective. It does not propose a single network architecture or training objective as a complete solution. Instead, it examines the structural trade-offs that arise when designing a multimodal music foundation model that jointly represents singing voice and score information. The paper considers how architectural decisions interact with data infrastructure, evaluation methodology, robustness, fairness, deployment, and policy. In doing so, it contributes a conceptual framework for researchers and engineers who must build systems that operate beyond narrow benchmark settings. The focus is therefore on system-level reasoning rather than isolated algorithmic contributions.

2. Conceptual Foundations and Multimodal Representation Challenges

Music modeling has historically developed along two separate paths. Symbolic music modeling has concentrated on discrete representations such as piano rolls, MIDI, and music notation, enabling long-range structural modeling and controllable generation. The Music Transformer, for example, demonstrated that self-attention mechanisms can capture long-term dependencies in symbolic sequences [5]. In parallel, audio-centric modeling has focused on waveform or spectrogram inputs, often leveraging large datasets of recorded performances. The MAESTRO dataset enabled factorized piano music modeling by aligning MIDI and audio, but its focus was primarily instrumental piano rather than singing [6]. Speech-oriented self-supervised models such as wav2vec 2.0 have shown that robust audio representations can be learned from raw speech, providing a useful reference for singing voice encoders [7]. Signal processing libraries such as librosa have long supported feature extraction and analysis, but they operate below the level of semantic score alignment [8].

The core representation challenge is that audio and score occupy different regions of a complex modality space. Symbolic scores are compact, editable, and semantically interpretable, but they omit many expressive dimensions. Singing audio is information-dense and acoustically rich, but it is entangled with speaker identity, recording conditions, language, and performance style. A joint representation should preserve score-level controllability while capturing performance-level nuance. This requires more than aligning two embedding distributions. It requires modeling the conditional relationship between discrete musical intentions and continuous acoustic realizations.

Another challenge is temporal granularity. Score events are discrete and often sparse, while audio frames are dense and continuous. Any joint representation must therefore learn an alignment that is robust to variable note durations, ornamentation, and phrase-level timing. The alignment problem becomes especially difficult for singing because the voice is also a linguistic channel. Phonemes, syllables, and word boundaries introduce additional structure that is not explicitly present in most music scores. A system that ignores linguistic content may fail to distinguish between different vowels sung on the same pitch, while a system that overweights linguistic structure may lose musical nuance.

These challenges imply that joint singing voice and score representation learning cannot be treated as a simple extension of text-image contrastive learning. The design space is broader and includes choices about temporal resolution, modality-specific encoders, cross-modal attention, latent regularization, and decoding heads. Each choice carries implications for computational cost, data requirements, interpretability, and downstream utility.

3. Architectural Design and System-Level Trade-offs

A multimodal music foundation model for singing voice and score typically requires at least three architectural components: a singing voice encoder, a symbolic score encoder, and an alignment mechanism that links their representations. Two-tower architectures encode each modality independently and align them in a shared latent space using a contrastive objective. This design is computationally efficient because the two encoders can be trained separately and the alignment layer remains light. However, two-tower models often struggle with fine-grained temporal correspondence because the interaction between modalities is restricted to a global or segment-level similarity score.

Cross-attention architectures offer richer interaction by allowing score embeddings to attend directly to audio embeddings at multiple temporal scales. This supports finer alignment and can improve tasks such as note-level transcription or score-informed source separation. However, cross-attention is more memory-intensive and requires careful design to avoid overfitting to spurious correlations. It also complicates retrieval because the encoders are no longer independently embeddable, which may be a drawback for large-scale indexing and search.

A third design family uses sequence-to-sequence decoders to reconstruct one modality from the other. For example, a score encoder may produce a conditioning representation that guides an audio decoder, or an audio encoder may produce features that are decoded into score tokens. This approach is attractive for generative tasks but can be less suitable for discriminative tasks such as retrieval or classification. The choice of architecture therefore depends on the intended ecosystem of downstream applications. A foundation model intended for music education may prioritize interpretable score-to-audio generation, while a retrieval system may benefit from a two-tower design with efficient embedding lookup.

System-level trade-offs also arise in temporal modeling. Convolutional encoders provide efficient local feature extraction but may miss long-range musical structure. Transformer encoders capture long-range dependencies but require more data and compute. Hybrid architectures that combine convolutional front-ends with transformer backbones offer a pragmatic compromise. In singing voice analysis, the front-end must handle pitch, timbre, and phoneme information, while the backbone should model phrase-level dynamics. The score encoder faces a different challenge: music scores contain hierarchical structure that is not always linear in time. Repeats, codas, and multi-voice textures introduce graph-like dependencies that a purely sequential model may fail to capture.

The alignment mechanism itself must balance local precision and global coherence. Some systems align at the note level, while others align at the frame or beat level. Note-level alignment is interpretable but may miss expressive micro-timing. Frame-level alignment captures fine detail but can be noisy and computationally expensive. A robust foundation model may need multiple alignment heads operating at different granularities, with careful multi-task weighting to prevent one objective from dominating training.

4. Data Infrastructure and Governance

Data infrastructure is a decisive but often underestimated factor in multimodal music modeling. Building a joint singing voice and score foundation model requires paired audio and symbolic score data with reliable alignment. A recently introduced global multi-technique singing corpus with realistic music scores provides a valuable resource for this purpose because it spans multiple singing techniques and offers realistic score annotations [9]. However, no single corpus can cover the full diversity of musical traditions, languages, vocal styles, and recording conditions. Data infrastructure must therefore be designed as an evolving ecosystem of datasets, annotation pipelines, and quality control mechanisms.

Paired data collection presents several governance challenges. Singing voice recordings often involve performers whose consent and rights must be respected. Symbolic scores may be protected by copyright, and derivative annotations may introduce additional licensing complexity. A responsible data infrastructure should document provenance, licensing terms, and permitted use cases for every source. It should also separate training, validation, and test splits in ways that prevent leakage across singers or compositions. Leakage can inflate evaluation scores and undermine claims of generalization.

The temporal alignment between audio and score is rarely perfect in real-world data. Annotation noise, performance deviations, and score ambiguities require robust preprocessing and alignment algorithms. Human verification is expensive but necessary for high-quality benchmarks. Automated alignment can scale data collection but may introduce systematic errors that are later learned by the model. Governance mechanisms should track annotation confidence and support targeted human review where uncertainty is high.

Institutional and community governance also matters. Music data often crosses national and cultural boundaries. Different jurisdictions have different rules about personal data, copyright, and moral rights. A foundation model trained on global singing data must account for these differences rather than assuming a single legal regime. Dataset documentation practices from natural language processing and computer vision can be adapted, but music-specific categories such as performer identity, recording context, and score arrangement need to be added. These governance choices are not external constraints; they shape the representational capacity and social legitimacy of the resulting model.

5. Training Strategy and Evaluation Methodology

Training a joint singing voice and score representation model requires careful orchestration of multiple objectives. The most common approach is contrastive learning, in which paired audio and score segments are pulled together while unpaired segments are pushed apart. Negative sampling is critical. Random negatives are easy to obtain but may be too easy, leading to representations that do not capture fine-grained musical structure. Hard negative mining can improve discrimination but increases computational cost and may introduce biases if the hard negatives are not representative of real confusions. Curriculum learning can gradually increase difficulty, beginning with clean instrumental-aligned data and progressing to noisy singing performances.

Evaluation is equally challenging. Common music information retrieval metrics provide useful but partial signals. The `mir_eval` library offers transparent implementations of transcription, alignment, and source separation metrics [10]. However, these metrics often assume deterministic score-to-audio correspondences and may penalize legitimate expressive variation. Expressive performance modeling research has shown that the same score can be realized in many musically valid ways [11]. Cross-domain pretraining strategies, such as those explored in LakhNES, suggest that symbolic and audio models can benefit from shared representations even when paired data are limited [12]. Discriminative tasks such as music classification provide additional probes for representation quality, but they do not directly measure alignment [13]. Pitch estimation methods [14] and perceptual resynthesis measures [15] can complement objective metrics, but they require careful interpretation.

A robust evaluation framework should include multiple task families: retrieval, transcription, generation, source separation, and human perceptual judgment. Retrieval tasks measure whether the representation places related audio and score segments close together. Transcription tasks measure whether the model can recover score events from audio. Generation tasks measure whether the model can produce plausible singing audio from score conditioning. Human evaluations are particularly important for singing because perceptual quality is not fully captured by spectral or note-level metrics. Listening tests should be conducted with diverse listeners and should include controls for preference, naturalness, and score fidelity.

6. Robustness and Failure Modes

Robustness is a central concern for multimodal music systems because real-world conditions differ sharply from curated training data. Singing voice recordings may include background noise, reverberation, microphone variation, and codec artifacts. Scores may contain ambiguous ornaments, incomplete dynamics, or unconventional notation. A model that performs well on studio-quality data may fail when exposed to amateur recordings or live performances. Distribution shift can be gradual, as in changing recording technology, or abrupt, as in a new singing style.

The relationship between score and performance can also vary across genres. In classical singing, adherence to score may be relatively strict, while in jazz or popular music, improvisation and ornamentation are expected. A model trained predominantly on one tradition may penalize legitimate deviations in another. Robustness therefore requires explicit modeling of stylistic variation rather than treating deviation as noise. This is a structural property of the training objective and data distribution, not merely a regularization detail.

Failure modes in multimodal learning often stem from modality imbalance. If the audio encoder is much stronger than the score encoder, the model may ignore symbolic information and rely on acoustic cues alone. Conversely, if the score encoder dominates, the model may fail to capture expressive performance. Monitoring modality utilization is therefore important during training and evaluation. Diagnosis can involve probing the latent space with controlled perturbations, but interpretability methods remain limited for music. Research on machine learning in music creation has emphasized that algorithmic outputs can be technically impressive while musically or culturally inappropriate [16]. Similar concerns apply to representation learning.

Large models may also exhibit hallucination or overgeneration, producing plausible but incorrect alignments. This is connected to broader concerns about stochastic parrots in language models, where fluency can mask a lack of grounding [17]. In music, grounding is inherently multimodal, and alignment errors may be harder to detect than textual errors. Robustness work should therefore include targeted stress tests for pitch, rhythm, phonetic content, and singer identity.

7. Fairness, Interpretability, and Accountability

Fairness in music foundation models is often discussed in terms of representation. A model trained primarily on Western classical or popular music may underrepresent non-Western scales, vocal techniques, and notational systems. Singing voice data also intersects with language, dialect, gender, and age. If the training data are dominated by a narrow demographic, the resulting representations may perform poorly for underrepresented groups. Fairness therefore depends on data curation, but also on evaluation protocols that measure performance across subpopulations and musical traditions.

Interpretability is a prerequisite for meaningful fairness audits. Model cards provide a structured way to document intended use, training data, performance characteristics, and limitations [18]. In music, model cards could include information about musical genres, languages, recording conditions, and vocal characteristics. Attribution methods such as Grad-CAM offer visual explanations for saliency in convolutional networks, but their adaptation to audio and symbolic score inputs requires care [19]. Dataset documentation tools developed in natural language processing can also support transparency in music corpora [20]. These instruments help external stakeholders understand what a model has learned and where it may fail.

Accountability extends beyond technical documentation. Organizations that deploy music foundation models should establish processes for contestability, redress, and continuous monitoring. If a singing voice model is used in education, for example, a student should be able to challenge incorrect feedback generated by the model. If it is used in creative production, performers should have clarity about attribution and ownership. These mechanisms are often absent from research prototypes but become essential in deployed systems.

8. Deployment, Sustainability, and Policy Implications

Deployment of multimodal music foundation models raises practical questions about latency, memory, and energy consumption. Large transformer models are expensive to serve, especially when they must process high-resolution audio. Singing voice and score alignment may require multiple passes over temporal features, increasing computational burden. Techniques such as knowledge distillation, quantization, and pruning can reduce costs, but

they may also degrade alignment quality. System designers must choose operating points that balance performance and resource use. Cross-modal generative models have demonstrated that large-scale pretraining can enable zero-shot capabilities, but these gains come with substantial training cost [21].

Sustainability is an increasingly important consideration. The carbon emissions associated with large model training have been documented, and music models are no exception [22]. Training on long audio sequences and large multilingual singing corpora can be particularly energy-intensive. Sustainability strategies include using pretrained audio encoders rather than training from scratch, selecting efficient architectures, and reporting energy consumption alongside accuracy. Green AI principles should be integrated into music modeling practice rather than treated as an afterthought. Audio-language contrastive pretraining can also offer efficiency gains by reusing existing multimodal alignment techniques [23], but domain adaptation for singing remains necessary.

Policy implications span copyright, cultural heritage, and performer rights. Generative music models may produce outputs that resemble existing works, raising questions about infringement and fair use. Score-conditioned singing models may reproduce a performer’s timbre or style, implicating rights of publicity and moral rights. Spectrogram diffusion models can generate realistic instrument and voice performances, making provenance tracking more urgent [24]. Policymakers and system builders should collaborate on watermarking, provenance metadata, and licensing frameworks that accommodate both creative innovation and rights protection.

The governance of music foundation models should also consider access. If high-quality singing voice and score models are controlled by a small number of organizations, independent researchers, educators, and cultural institutions may be excluded. Open model releases, community datasets, and transparent evaluation benchmarks can mitigate concentration of power. At the same time, openness must be balanced against the risk of misuse, including unauthorized voice cloning and cultural appropriation. These trade-offs are not purely technical; they require institutional design, stakeholder engagement, and iterative policy development.

9. Conclusion

Joint representation learning for singing voice and music scores is a foundational problem for the next generation of multimodal music systems. The alignment of continuous vocal audio and symbolic score information requires more than a powerful network. It demands a coherent system architecture, robust data governance, carefully designed training and evaluation protocols, and explicit attention to fairness, interpretability, deployment, and sustainability. This paper has examined these dimensions in an integrated way, arguing that technical performance and socio-technical responsibility are mutually constitutive rather than separable concerns.

The architectural trade-offs between two-tower contrastive designs, cross-attention mechanisms, and sequence-to-sequence decoders reflect different priorities for retrieval, generation, transcription, and interpretability. Data infrastructure must support diverse musical traditions and singing styles while respecting performer rights and legal constraints. Training strategies must balance local alignment precision with global musical coherence. Evaluation must go beyond narrow metrics to include perceptual judgment and stress testing. Robustness requires explicit modeling of stylistic variation and modality balance. Fairness

and accountability require documentation, interpretability, and contestability. Policy must address copyright, voice rights, provenance, and access.

Future work should develop open multimodal music benchmarks that capture diverse languages, singing techniques, and score formats. It should also investigate lightweight architectures that enable sustainable deployment while preserving alignment quality. Interdisciplinary collaboration among engineers, musicologists, legal scholars, and policy experts is essential. By treating joint singing voice and score representation learning as an infrastructure problem, the field can build music foundation models that are not only technically impressive but also socially grounded and broadly useful.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748-8763.
3. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). MusicLM: Generating music from text. arXiv preprint arXiv:2301.11325.
4. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Defossez, A. (2024). Simple and controllable music generation. Advances in Neural Information Processing Systems, 36.
5. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music Transformer: Generating music with long-term structure. arXiv preprint arXiv:1809.04281.
6. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J., & Eck, D. (2019). Enabling factorized piano music modeling with the MAESTRO dataset. International Conference on Learning Representations.
7. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.
8. McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference, 18-25.
9. Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... & Zhao, Z. (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. Advances in Neural Information Processing Systems, 37, 1117-1140.
10. Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. W. (2014). mir_eval: A transparent implementation of common MIR metrics. Proceedings of the 15th International Society for Music Information Retrieval Conference, 367-372.

11. Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32(4), 955-967.
12. Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W., & McAuley, J. (2019). LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 685-692.
13. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2392-2396.
14. Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2017). Deep salience representations for f0 estimation in polyphonic music. *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 63-70.
15. Simonetta, F., Ntalampiras, S., & Avanzini, F. (2019). A perceptual measure for evaluating the resynthesis of automatic music transcriptions. *Multimedia Tools and Applications*, 78(19), 27745-27764.
16. Sturm, B. L., Ben-Tal, O., Monaghan, U., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E., & Pachet, F. (2019). Machine learning research that matters for music creation: A case study. *Journal of New Music Research*, 48(1), 36-55.
17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
19. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618-626.
20. Lhoest, Q., Villanova del Moral, A., Jernite, Y., Ramesh, A., et al. (2021). Datasets: A community library for natural language processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 175-184.
21. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning*, 8821-8831.
22. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
23. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*.

24. Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., & Engel, J. (2022). Multi-instrument music synthesis with spectrogram diffusion. Proceedings of the 23rd International Society for Music Information Retrieval Conference, 598-606.