CURRENT RESEARCH JOURNAL OF PEDAGOGICS (ISSN: 2767-3278)

VOLUME: Vol.07 Issue05 2026
Page: - 01-09

RESEARCH ARTICLE OPEN ACCESS

Advanced Conditional Frameworks for Probabilistic Sound Generation Enabling Greater Authenticity and Tonal Regulation

Dr. Budi Santoso

Department of Computer Science, Universitas Gadjah Mada, Yogyakarta, Indonesia

Received: 08 March 2026 **Accepted:** 05 April 2026 Published**:** 01 May 2026

ABSTRACT

Probabilistic sound generation has undergone a transformative evolution with the emergence of deep generative models, particularly diffusion-based architectures, variational autoencoders, and generative adversarial networks. While these approaches have significantly improved the realism of synthesized audio, they often suffer from limitations in controllability and tonal precision. This paper investigates advanced conditional frameworks designed to enhance both authenticity and fine-grained acoustic regulation in generative audio systems. By synthesizing insights from foundational generative modeling techniques and recent advancements in diffusion-based sound synthesis, this study proposes a structured analytical perspective on multi-condition integration strategies.

The research explores how conditioning mechanisms—such as textual prompts, spectral features, symbolic representations, and performance parameters—affect the probabilistic modeling of sound. It further evaluates the interplay between conditioning modalities and generative architectures, highlighting how diffusion models enable iterative refinement processes that align outputs with desired tonal characteristics. Theoretical grounding is provided through probabilistic modeling frameworks, including latent variable models and score-based generative processes, enabling a deeper understanding of how conditional signals influence output distributions.

A critical component of the study is the comparative analysis of conditioning strategies across architectures, including waveform-based synthesis (e.g., WaveNet), spectrogram-based modeling, and latent diffusion systems. The paper identifies key challenges such as mode collapse, over-conditioning, and loss of diversity, and examines mitigation strategies through hierarchical conditioning and adaptive weighting schemes. Additionally, evaluation metrics such as Fréchet Audio Distance and perceptual realism measures are analyzed to assess improvements in generated audio quality.

The findings suggest that advanced conditional frameworks significantly enhance both perceptual realism and controllability, particularly when multi-modal conditioning is incorporated. However, trade-offs emerge between flexibility and computational complexity, necessitating optimized architectures. This work contributes to the field by offering a comprehensive framework for understanding conditional sound generation and outlining future directions for scalable, interpretable, and high-fidelity audio synthesis systems.

Keywords: Probabilistic Sound Generation, Diffusion Models, Conditional Generative Models, Audio Synthesis, Tonal Control, Neural Audio Processing, Multi-Modal Conditioning, Latent Diffusion, Generative Adversarial Networks.

INTRODUCTION

The synthesis of realistic and controllable audio has long been a central objective in computational music generation and speech processing. Early approaches to sound synthesis, including physical modeling and concatenative techniques, focused primarily on replicating acoustic phenomena through deterministic or rule-based systems (Smith, 1992; Schwarz, 2005). While these methods provided high levels of interpretability and physical accuracy, they were constrained by limited flexibility and scalability. The advent of neural generative models introduced a paradigm shift, enabling data-driven approaches capable of capturing complex temporal and spectral dependencies in audio signals.

Generative models such as WaveNet demonstrated the feasibility of producing high-fidelity raw audio through autoregressive modeling (van den Oord, 2016). Subsequent developments in variational autoencoders (Kingma and Welling, 2014) and generative adversarial networks (Goodfellow, 2020) expanded the design space for probabilistic sound generation, offering new mechanisms for learning latent representations and generating diverse outputs. However, these models often lacked precise control over generated content, particularly in musical contexts where tonal structure, timbre, and expressive dynamics are critical.

Recent advancements in diffusion-based generative models have addressed several of these limitations by introducing iterative denoising processes that progressively refine generated samples (Ho et al., 2020). These models have demonstrated remarkable success in both image and audio domains, enabling high-quality synthesis with improved stability. In the context of sound generation, diffusion models such as Diff-TTS and spectrogram-based diffusion systems have shown the ability to produce realistic audio conditioned on textual or symbolic inputs (Jeong et al., 2021; Hawthorne, 2022). Despite these advances, achieving fine-grained tonal control remains a significant challenge.

The core problem addressed in this paper is the integration of advanced conditional frameworks that enable both authenticity and precise tonal regulation in probabilistic sound generation systems. Conditioning mechanisms play a crucial role in guiding generative models toward desired outputs, yet their design and implementation vary widely across architectures. Existing approaches often rely on single-modal conditioning, limiting their ability to capture the multi-dimensional nature of musical and acoustic expression.

This research aims to systematically analyze and extend conditional frameworks by examining multi-modal conditioning strategies, hierarchical representations, and adaptive conditioning mechanisms. The study is motivated by the need for more expressive and controllable generative systems capable of supporting applications such as music composition, virtual instrument design, and interactive audio synthesis.

The significance of this work lies in its integration of theoretical and practical perspectives. By bridging probabilistic modeling theory with empirical advancements in neural audio synthesis, the paper provides a comprehensive framework for understanding how conditional signals influence generative processes. Furthermore, it identifies key challenges and proposes potential solutions, contributing to the development of next-generation audio generation systems.

LITERATURE REVIEW

The evolution of probabilistic sound generation is deeply rooted in both traditional signal processing techniques and modern machine learning approaches. Early work in digital sound synthesis, such as physical modeling using digital waveguides, emphasized the simulation of acoustic systems through mathematical representations (Smith, 1992). Similarly, concatenative synthesis leveraged recorded audio segments to construct new sounds, offering high realism but limited generative flexibility (Schwarz, 2006; Sturm, 2006). These approaches laid the groundwork for understanding sound as a structured, manipulable signal.

The introduction of neural generative models marked a significant departure from deterministic synthesis. Variational autoencoders provided a probabilistic framework for learning latent representations of audio data, enabling interpolation and controlled generation (Kingma and Welling, 2014). Generative adversarial networks further advanced the field by introducing adversarial training mechanisms that improved the realism of generated outputs (Goodfellow, 2020). However, GAN-based approaches often struggled with stability and mode collapse, particularly in high-dimensional audio domains.

Waveform-based models such as WaveNet demonstrated

the potential of autoregressive architectures to generate high-fidelity audio by modeling temporal dependencies at the sample level (van den Oord, 2016). Subsequent work extended these models to incorporate conditioning signals, enabling applications such as text-to-speech synthesis and musical timbre control (Engel, 2017; Kim et al., 2019). Despite their success, autoregressive models are computationally expensive and lack scalability for long-duration audio generation.

Diffusion models have emerged as a powerful alternative, offering improved stability and quality through iterative refinement processes. The foundational work on denoising diffusion probabilistic models established a theoretical framework for generating data by reversing a stochastic noise process (Ho et al., 2020). In the audio domain, diffusion-based approaches have been applied to text-to-speech synthesis, music generation, and spectrogram modeling (Jeong et al., 2021; Huang, 2023). These models benefit from their ability to incorporate conditioning signals at multiple stages of the generation process.

Recent research has focused on enhancing controllability through advanced conditioning mechanisms. Multi-modal conditioning, which integrates textual, symbolic, and acoustic inputs, has been shown to improve the expressiveness of generated audio (Agostinelli, 2023; Schneider, 2023). Techniques such as feature-wise linear modulation (Perez et al., 2018) and hierarchical conditioning structures enable more precise control over output characteristics. Additionally, latent diffusion models have introduced efficient representations that reduce computational complexity while maintaining high-quality synthesis (Rombach et al., 2022).

Evaluation of generative audio models remains a complex challenge. Metrics such as Fréchet Audio Distance provide quantitative measures of similarity between generated and real audio distributions, while perceptual evaluations assess subjective quality (Kilgour et al., 2019). These evaluation frameworks highlight the trade-offs between realism, diversity, and controllability, underscoring the need for balanced model design.

Despite significant progress, several research gaps persist. Existing models often struggle to balance multi-modal conditioning with computational efficiency, and the integration of hierarchical control mechanisms remains underexplored. Furthermore, the relationship between conditioning strength and output diversity is not fully understood, limiting the practical applicability of these systems.

ADVANCED CONDITIONAL FRAMEWORK DESIGN FOR PROBABILISTIC SOUND GENERATION

The design of advanced conditional frameworks is central to improving the controllability, realism, and expressiveness of probabilistic sound generation systems. Traditional generative models often rely on single-condition inputs, which limits their ability to capture the complex, multi-dimensional nature of audio signals. In contrast, modern frameworks integrate multiple conditioning mechanisms that operate across different representational levels, enabling more nuanced and flexible generation.

These frameworks are grounded in probabilistic modeling principles, where conditional distributions guide the generation process. By incorporating structured inputs such as textual descriptions, symbolic representations, and acoustic features, conditional systems can align generated outputs with desired attributes. The integration of these diverse inputs requires sophisticated architectural designs, including attention mechanisms, hierarchical encoders, and latent variable models.

From a functional perspective, advanced conditional frameworks must address three core challenges: representation alignment, conditioning integration, and output consistency. Representation alignment ensures that inputs from different modalities are mapped into compatible feature spaces. Conditioning integration involves effectively combining multiple signals without introducing conflicts. Output consistency requires maintaining coherence across temporal and spectral dimensions of generated audio.

The following subsections examine key conditioning strategies that underpin modern probabilistic sound generation systems.

1 Multi-Aspect Conditioning Strategies

Multi-aspect conditioning involves the simultaneous use of multiple input signals to guide the generative process. These inputs may include textual descriptions, musical scores, timbral embeddings, and performance parameters. The objective is to capture different dimensions of audio

characteristics, such as semantic meaning, structural composition, and acoustic properties.

Diffusion-based models such as those proposed by Ho et al. (2020) incorporate conditioning through noise prediction networks, where auxiliary inputs influence each denoising step. This allows the model to maintain alignment with conditioning signals throughout the generation process. Similarly, text-to-music systems like MusicLM (Agostinelli, 2023) and Noise2Music (Huang, 2023) demonstrate how semantic conditioning can drive high-level musical structure.

A key advantage of multi-aspect conditioning is enhanced controllability. By adjusting individual conditioning inputs, users can manipulate specific attributes of the generated output without affecting others. For example, altering a textual prompt may change the genre, while modifying a timbral embedding affects the instrument sound.

However, integrating multiple conditioning signals introduces challenges related to signal interference and prioritization. Attention-based mechanisms and gating functions are commonly used to dynamically weight different inputs, ensuring balanced influence across the generation process.

2 Cross-Modal and Semantic Conditioning

Cross-modal conditioning extends the generative framework by incorporating inputs from different data modalities, such as text, images, and symbolic representations. This approach leverages shared semantic spaces to enable coherent translation between modalities, significantly expanding the scope of generative applications.

Models inspired by contrastive learning techniques, such as those discussed by Radford (2021), map textual and audio inputs into aligned embedding spaces. This alignment enables the model to interpret semantic relationships and generate outputs that correspond to high-level descriptions. For instance, a textual prompt describing a “soft piano melody with ambient background” can be translated into a corresponding audio waveform.

In probabilistic sound generation, cross-modal conditioning is particularly effective when combined with diffusion models. Latent diffusion frameworks (Rombach et al., 2022) operate in compressed feature spaces, allowing efficient integration of multimodal inputs. Systems like Moûsai (Schneider et al., 2023) further demonstrate the ability to maintain long-term coherence in music generation through context-aware conditioning.

Despite its potential, cross-modal conditioning faces challenges related to semantic ambiguity and alignment errors. Differences in modality-specific representations can lead to inconsistencies between input intent and generated output. Addressing these issues requires robust embedding techniques and large-scale training data.

3 Performance-Aware and Temporal Conditioning Mechanisms

Performance-aware conditioning focuses on capturing dynamic aspects of audio signals, including timing, articulation, and expressive variations. Unlike static conditioning, which defines global attributes, performance-aware mechanisms operate over temporal sequences, enabling the generation of realistic and expressive audio.

Score-to-audio systems such as Performancenet (Wang and Yang, 2019) and Deep Performer (Dong et al., 2022) utilize symbolic representations to guide temporal evolution. These models incorporate recurrent or transformer-based architectures to model dependencies across time, ensuring continuity and coherence in generated sequences.

Diffusion-based approaches further enhance temporal conditioning by integrating time-dependent noise schedules. Models like Diff-TTS (Jeong et al., 2021) and spectrogram diffusion systems (Hawthorne, 2022) maintain temporal consistency through iterative refinement, allowing precise control over rhythm and dynamics.

Performance-aware conditioning also enables real-time interaction, where users can influence generation through live inputs such as MIDI controllers or gesture-based interfaces. This capability is particularly valuable in creative applications, where responsiveness and expressiveness are critical.

However, modeling temporal dependencies introduces computational complexity and requires large datasets to capture diverse performance styles. Additionally, ensuring

synchronization between temporal and spectral features remains a significant challenge in complex audio generation tasks

4 Hierarchical Multi-Level Conditioning Architectures

Hierarchical conditioning introduces structured control across multiple abstraction layers, enabling fine-grained manipulation of musical attributes. Unlike flat conditioning, hierarchical frameworks decompose the generation process into macro, meso, and micro levels, each responsible for distinct musical properties such as composition, arrangement, and timbre synthesis.

At the macro level, high-level representations such as textual prompts or symbolic sequences (e.g., MIDI) guide the global structure of generated audio. Models like MusicLM (Agostinelli, 2023) and Noise2Music (Huang, 2023) utilize semantic embeddings derived from textual descriptions to influence compositional coherence. This level ensures that generated outputs align with thematic and stylistic expectations.

The meso level focuses on performance attributes, including rhythm, tempo, and articulation. Systems such as Performancenet (Wang and Yang, 2019) and Deep Performer (Dong et al., 2022) integrate score-based inputs with expressive performance modeling, enabling dynamic variation within structured compositions. Conditioning at this level enhances realism by capturing human-like variations.

At the micro level, signal-level conditioning governs timbral characteristics and acoustic fidelity. Techniques such as WaveNet-based synthesis (van den Oord, 2016) and DDSP-inspired approaches (Wu, 2022) operate directly on waveform or spectral representations. Diffusion-based models further refine these signals through iterative denoising, allowing precise control over audio textures.

The integration of these hierarchical levels enables comprehensive control across the generative pipeline. However, challenges arise in maintaining consistency across levels, particularly when conflicting conditioning signals are introduced. Synchronization mechanisms and cross-level attention modules are therefore essential for coherent synthesis.

5 Latent Space Conditioning and Representation Learning

Latent space conditioning represents a critical advancement in probabilistic sound generation, enabling efficient manipulation of high-dimensional audio representations. Instead of directly conditioning on raw inputs, models encode information into compact latent vectors that capture essential semantic and acoustic features.

Variational Autoencoders (Kingma and Welling, 2014) provide a probabilistic framework for latent representation learning, where audio data is mapped to a continuous latent space. This facilitates smooth interpolation between musical attributes and supports generative diversity. Latent diffusion models (Rombach et al., 2022) extend this concept by performing diffusion processes in compressed latent spaces, significantly reducing computational complexity while preserving quality.

In music generation, latent conditioning allows for disentanglement of features such as pitch, timbre, and rhythm. For instance, MIDI-DDSP (Wu, 2022) separates control parameters for pitch and synthesis, enabling independent manipulation. Similarly, Moûsai (Schneider et al., 2023) leverages long-context latent representations to generate coherent musical sequences over extended durations.

Latent conditioning also supports cross-modal integration. Models can map textual, visual, or symbolic inputs into shared latent spaces, enabling multimodal generation. CLIP-inspired approaches (Radford, 2021) demonstrate how semantic alignment across modalities enhances conditioning effectiveness.

Despite its advantages, latent space conditioning introduces interpretability challenges. The abstract nature of latent variables makes it difficult to directly control specific attributes without extensive calibration. Furthermore, disentanglement is not guaranteed and often requires additional regularization techniques.

6 Adaptive Conditioning and Real-Time Control Mechanisms

Adaptive conditioning frameworks enable dynamic modification of generative processes during inference, supporting interactive and real-time applications. Unlike static conditioning, adaptive systems adjust conditioning parameters based on feedback, user input, or environmental changes.

One approach involves reinforcement learning-based control, where the model iteratively refines outputs to optimize predefined objectives such as realism or stylistic accuracy. Another method employs attention-based modulation, allowing the model to dynamically prioritize different conditioning signals at various stages of generation.

Real-time systems such as interactive music synthesis platforms benefit significantly from adaptive conditioning. For example, performance-conditioned diffusion models (Maman et al., 2024) allow users to influence output characteristics during generation, enabling applications in live music production and digital instrument design.

Adaptive conditioning also supports personalization, where models learn user preferences over time and adjust outputs accordingly. This is particularly relevant in creative domains, where subjective criteria such as aesthetic quality and emotional resonance play a crucial role.

However, adaptive systems introduce additional complexity in terms of stability and computational overhead. Ensuring real-time responsiveness while maintaining high-quality output remains a significant challenge. Moreover, feedback-driven adaptation may lead to unintended biases or overfitting to specific user preferences.

ANALYTICAL EVALUATION OF CONDITIONING STRATEGIES IN SOUND GENERATION

The evaluation of conditional frameworks in probabilistic sound generation requires a multi-dimensional approach, encompassing objective metrics, subjective assessments, and robustness analysis. Given the complexity of audio perception, no single metric can fully capture the quality and realism of generated outputs.

Objective evaluation metrics such as Fréchet Audio Distance (FAD) (Kilgour et al., 2019) provide quantitative measures of similarity between generated and real audio distributions. These metrics are particularly useful for benchmarking model performance across datasets. However, they often fail to capture perceptual nuances such as emotional expressiveness and stylistic coherence.

Subjective evaluation remains essential for assessing perceptual quality. Human listening tests, including mean opinion scores (MOS), provide insights into realism and user satisfaction. Studies have shown that diffusion-based models generally outperform GAN-based approaches in subjective evaluations due to their ability to produce smoother and more coherent audio (Ho et al., 2020; Jeong et al., 2021).

Comparative analysis reveals that multi-aspect conditioning significantly enhances both objective and subjective performance. Models incorporating hierarchical and latent conditioning achieve higher fidelity and greater controllability compared to single-condition systems. For instance, spectrogram diffusion models (Hawthorne, 2022) demonstrate improved multi-instrument synthesis through integrated conditioning mechanisms.

Robustness evaluation examines the model's ability to maintain performance under varying conditions, such as noisy inputs or incomplete conditioning signals. Diffusion models exhibit strong robustness due to their iterative refinement process, which mitigates the impact of noise. However, they are computationally intensive, limiting scalability.

A critical limitation across evaluation methods is the lack of standardized benchmarks for conditional sound generation. Existing datasets often fail to capture the diversity of real-world audio scenarios, leading to potential overfitting. Future research must focus on developing comprehensive evaluation frameworks that integrate objective metrics with perceptual and contextual assessments.

RESULTS

The analytical investigation of advanced conditional frameworks for probabilistic sound generation reveals several key findings regarding system performance, controllability, and perceptual quality.

First, diffusion-based architectures consistently demonstrate superior performance in generating high-fidelity audio compared to GAN and VAE-based models. Their iterative denoising process enables gradual refinement of audio signals, resulting in smoother and more realistic outputs. This advantage is particularly evident in tasks requiring detailed acoustic modeling, such as multi-instrument synthesis and expressive speech generation (Ho et al., 2020; Jeong et al., 2021).

Second, the integration of multi-aspect conditioning significantly enhances model controllability. Systems that incorporate hierarchical, latent, and performance-based conditioning achieve greater flexibility in manipulating musical attributes. This allows for precise control over timbre, rhythm, and structure, enabling more diverse and contextually appropriate outputs (Maman et al., 2024; Wu, 2022).

Third, latent space conditioning emerges as a critical factor in improving computational efficiency without compromising quality. By operating in compressed representation spaces, models reduce processing requirements while maintaining expressive capabilities. This is particularly beneficial for large-scale applications and real-time systems (Rombach et al., 2022).

Fourth, cross-modal conditioning frameworks enable seamless integration of textual, symbolic, and audio inputs, expanding the scope of generative applications. Models such as MusicLM and Noise2Music demonstrate the feasibility of generating coherent musical compositions from textual descriptions, highlighting the potential for multimodal creativity (Agostinelli, 2023; Huang, 2023).

However, the findings also identify significant challenges. Computational complexity remains a major limitation, particularly for diffusion-based models, which require extensive iterative processing. Additionally, the lack of standardized evaluation metrics complicates performance comparison across models.

Overall, the results indicate that while advanced conditional frameworks substantially improve sound generation capabilities, further optimization is required to address scalability and evaluation challenges.

DISCUSSION

The findings underscore the transformative impact of advanced conditional frameworks on probabilistic sound generation, while also revealing critical theoretical and practical implications.

From a theoretical perspective, the success of multi-aspect conditioning highlights the importance of structured representation in generative modeling. The integration of hierarchical and latent conditioning aligns with broader trends in machine learning, where modular architectures enable more effective learning and control. This suggests that future research should focus on developing unified frameworks that seamlessly integrate multiple conditioning strategies.

Practically, the enhanced controllability offered by advanced frameworks has significant implications for creative industries. Applications in music production, film scoring, and interactive media can benefit from the ability to generate customized audio content with high precision. Real-time adaptive systems further expand these possibilities, enabling dynamic interaction between users and generative models.

However, the discussion must also address inherent trade-offs. While diffusion models offer superior quality, their computational demands limit their accessibility and scalability. This creates a tension between performance and efficiency, necessitating the development of optimized architectures and hardware acceleration techniques.

Another critical issue is the interpretability of conditional models. As conditioning mechanisms become more complex, understanding the relationship between inputs and outputs becomes increasingly challenging. This limits the ability to diagnose errors and refine model behavior, particularly in professional applications requiring precise control.

Ethical considerations also emerge, particularly in the context of content authenticity and ownership. The ability to generate highly realistic audio raises concerns about misuse, including deepfake audio and unauthorized replication of artistic styles. Addressing these issues requires the development of robust governance frameworks and detection mechanisms.

In comparison with existing literature, the findings align with prior studies emphasizing the superiority of diffusion models and the importance of conditioning in generative tasks (Ho et al., 2020; Rombach et al., 2022). However, this study extends the analysis by systematically examining the interplay between different conditioning strategies and their combined impact on performance.

CONCLUSION

This research provides a comprehensive analysis of advanced conditional frameworks for probabilistic sound generation, emphasizing their role in enhancing realism and tonal control. By integrating hierarchical, latent, and

adaptive conditioning mechanisms, modern generative models achieve unprecedented levels of fidelity and flexibility.

The study demonstrates that diffusion-based architectures, when combined with multi-aspect conditioning, represent the current state-of-the-art in audio synthesis. These systems enable precise manipulation of musical attributes, support cross-modal generation, and maintain robustness under varying conditions.

Despite these advancements, significant challenges remain. Computational complexity, lack of standardized evaluation frameworks, and interpretability issues limit the widespread adoption of these technologies. Addressing these challenges will require interdisciplinary efforts, combining advances in machine learning, signal processing, and human-computer interaction.

Future research should focus on developing efficient architectures, improving latent representation interpretability, and establishing comprehensive evaluation benchmarks. Additionally, ethical considerations must be integrated into system design to ensure responsible use of generative technologies.

In conclusion, advanced conditional frameworks represent a critical step toward achieving fully controllable and realistic sound generation systems, with far-reaching implications across scientific, industrial, and creative domains.

REFERENCES

1. Agostinelli, "MusicLM: Generating music from text," 2023, arXiv:2301.11325.

2. H. Dong, C. Zhou, T. Berg-Kirkpatrick, and J. J. McAuley, "Deep performer: Score-to-audio music performance synthesis," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Singapore, 2022 pp. 951–955.

3. J. H. Engel, "Neural audio synthesis of musical notes with wavenet autoencoders," in Proc. Int. Conf. Mach. Learn., Sydney, Australia, 2017, vol. 70, pp. 1068–1077.

4. J. Goodfellow, "Generative adversarial networks," Commun. ACM, vol. 63, pp. 139–144, 2020.

5. Hawthorne, "Multi-instrument music synthesis with spectrogram diffusion," in Proc. Int. Soc. Music Inf. Retrieval Conf., 2022, pp. 598–607.

6. Hawthorne, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in Proc. Int. Conf. Learn. Representations, New Orleans, Louisiana, USA, 2019.

7. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 6840–6851.

8. Q. Huang, "Noise2Music: Text-conditioned music generation with diffusion models," 2023, arXiv:2302.03917.

9. K. Karplus and A. Strong, "Digital synthesis of plucked-string and drum timbres," Comput. Music J., vol. 7, no. 2, pp. 43–55, 1983.

10. T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4401–4410.

11. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, 2020, pp. 8107–8116.

12. M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A denoising diffusion model for text-to-speech," in Proc. 22nd Annu. Conf. Int. Speech Commun. Assoc., Brno, Czechia, 2021, pp. 3605–3609.

13. J. W. Kim, R. M. Bittner, A. Kumar, and J. P. Bello, "Neural music synthesis for flexible timbre control," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Brighton, U.K., 2019, pp. 176–180.

14. P. Kingma and M. Welling, "Auto-encoding variational bayes," in Proc. 2nd Int. Conf. Learn. Representations, Y. Bengio and Y. LeCun, Eds. Banff, AB, Canada, Apr. 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

15. K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for

evaluating music enhancement algorithms," in Proc. Annu. Conf. Int. Speech Commun. Assoc., Graz, Austria, 2019, pp. 2350–2354.

16. H. Kim, S. Choi, and J. Nam, "Expressive acoustic guitar sound synthesis with an instrument-specific input representation and diffusion outpainting," in Proc. 2024 IEEE Int. Conf. Acoust., Speech Signal Process., 2024, pp. 7620–7624.

17. B. Maman, J. Zeitler, M. Müller, and A. H. Bermano, "Performance conditioning for diffusion-based multi-instrument music synthesis," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Seoul, South Korea, 2024, pp. 5045–5049.

18. B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in Proc. Int. Conf. Mach. Learn., Baltimore, Maryland, USA, 2022, pp. 14918–14934.

19. Maestre, R. Ramírez, S. Kersten, and X. Serra, "Expressive concatenative synthesis by reusing samples from real performance recordings," Comput. Music J., vol. 33, no. 4, pp. 23–42, 2009.

20. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, USA, 2022, pp. 10674–10685.

21. Radford, "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn., 2021, pp. 8748–8763.

22. Saharia, "Photorealistic text-to-image diffusion models with deep language understanding," in Proc. Adv. Neural Inf. Process. Syst. 35: Annu. Conf. Neural Inf. Process. Syst., New Orleans, LA, USA, 2022, pp. 36479–36494.

23. Schneider, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," 2023, arXiv:2301.11757.

24. Schwarz, "Current research in concatenative sound synthesis," in Proc. Int. Comput. Music Conf., Barcelona, Spain, Sep. 2005.

25. Schwarz, "Concatenative sound synthesis: The early years," J. New Music Reaserch, vol. 35, no. 1, pp. 3–22, Mar. 2006.

26. J. O. Smith, "Physical modeling using digital waveguides," Comput. Music J., vol. 16, no. 4, pp. 74–91, 1992.

27. B. L. Sturm, "Adaptive concatenative sound synthesis and its application to micromontage composition," Comput. Music J., vol. 30, no. 4, pp. 46–66, 2006.

28. J. Tseng, R. Castellon, and C. K. Liu, "Edge: Editable dance generation from music," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 448–458.

29. van den Oord, "WaveNet: A generative model for raw audio," in Proc. ISCA Speech Synth. Workshop, Sunnyvale, USA, 2016.

30. Wang and Y. Yang, "Performancenet: Score-to-audio music generation with multi-band convolutional residual network," in Proc. Conf. Artif. Intell., Honolulu, Hawaii, 2019, pp. 1174–1181.

31. Y. Wu, "MIDI-DDSP: Detailed control of musical performance via hierarchical modeling," in Proc. Int. Conf. Learn. Representations, 2022.

32. N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," IEEE/ACM Trans. Audio, Speech Lang. Process., vol. 30, pp. 495–507, 2022.