Improving Musical Accompaniment Co-creation via Diffusion Transformers

Javier Nistal[1] **Marco Pasini**[2*] **Stefan Lattner**[1]

[1]Sony Computer Science Laboratories, Paris
[2]Queen Mary University of London

Abstract

Building upon Diff-A-Riff, a latent diffusion model for musical instrument accompaniment generation, we present a series of improvements targeting quality, diversity, inference speed, and text-driven control. First, we upgrade the underlying autoencoder to a stereo-capable model with superior fidelity and replace the latent U-Net with a Diffusion Transformer. Additionally, we refine text prompting by training a cross-modality predictive network to translate text-derived CLAP embeddings to audio-derived CLAP embeddings. Finally, we improve inference speed by training the latent model using a consistency framework, achieving competitive quality with fewer denoising steps. Our model is evaluated against the original Diff-A-Riff variant using objective metrics in ablation experiments, demonstrating promising advancements in all targeted areas.

1 Introduction

Deep generative models for audio are rapidly gaining traction due to their potential to transform music creation and audio manipulation. Recent advancements [1]–[6] anticipate a future where humans and AI collaborate seamlessly to expand artistic expression. However, important challenges remain before these models can be fully integrated into professional music production workflows, including issues with audio quality, resource-intensive generation processes, and limited control mechanisms.

In response to these challenges, Diff-A-Riff [7] was recently introduced as a model specifically designed for music production purposes. Leveraging a Latent Diffusion Model [8] and a Consistency Autoencoder (CAE) [9], Diff-A-Riff can generate high-quality, pseudo-stereo, single-instrument accompaniments that adapt to different musical contexts—for example, generating a guitar track conditioned on a mixture of drums and bass. It also provides control through text prompts and audio style references using CLAP embeddings [10]. While Diff-A-Riff represents a significant step toward AI-assisted music co-creation tools, opportunities for improvement remain, particularly in achieving true stereo output with enhanced quality, computational efficiency, and better text-driven generation capabilities.

In this work, we present a series of enhancements to Diff-A-Riff, focusing on three key areas: quality, speed, and control. First, we enhance audio quality by upgrading the autoencoder to a stereo-capable model with improved fidelity [11] and transitioning from a latent diffusion U-Net to a Diffusion Transformer (DiT) architecture [12]. As shown in our experiments, these improvements generally lead to more diverse and higher-quality audio generation. Second, we refine the model's text-to-audio capabilities by training a separate diffusion model to mitigate the modality gap inherent in CLAP space [13], i.e., reducing the distance between audio-derived and text-derived CLAP embeddings. By

*This work is supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (EP/S022694/1) and Sony Computer Science Laboratories Paris.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

incorporating this model as an interface between text prompts and Diff-A-Riff, we experimentally show improved generation quality and a better resemblance of the audio with the intended prompts. Finally, to improve inference speed, we introduce consistency training to the latent model, maintaining competitive audio quality while enabling five-step generation. We rigorously evaluate our enhanced model against the original Diff-A-Riff using objective metrics in an ablation study. The results demonstrate improvements across the targeted areas, highlighting the potential of our improved model for practical music production applications.

2 Related Work

Autoregressive models trained on waveform samples [14], [15] produce high-fidelity audio but are computationally expensive. GANs [16] and VAEs [17] offer faster inference but struggle with long-term dependencies and complex musical structures [18], [19]. Denoising Diffusion Models have been applied to audio generation [20] but tend to be slow due to their iterative denoising process.

Recently, models trained on compressed audio representations via autoencoders have shown promise for efficiently modeling long-form music. Autoregressive models using discretized VQ-VAE representations [21] allow for handling long-term structures and multi-modal inputs [1], [2], [22]. Latent Diffusion Models, which operate in continuous latent spaces, achieve competitive quality at high sample rates [3]–[5]. Hybrid approaches [23] show promise for audio generation and editing.

Control mechanisms beyond text prompting are being explored to enhance precision in music production. Time-varying parameters [24], [25] and latent manipulations in text-audio spaces [26], [27] offer finer control, while inference-time optimization [28], [29] and guidance strategies [30] improve model control. Conditioning on audio signals has been effective for style transfer (e.g., melody or timbre [2], [31]) and accompaniment generation [7], [31]–[35]. Recent efforts combine music generation and source separation to create individual stems [36]–[39].

3 Background

3.1 Music2Latent2 (M2L2)[2] introduces an audio autoencoder designed to achieve high compression rates while preserving audio fidelity. M2L2 employs an autoregressive consistency model trained with causal masking to handle arbitrary audio lengths by processing the input in consecutive chunks. The model consists of three components: an encoder that extracts so-called *summary embeddings*– unordered latent representations capable of capturing global features of audio chunks– from the input complex Short-Time Fourier Transform (STFT) spectrogram, a decoder that produces upsampled audio features, and a consistency model that reconstructs the original spectrogram based on the features from the decoder. During inference, M2L2 uses a two-step decoding procedure, refining the generated audio by reintroducing noise into previously decoded segments. Compared to Music2Latent, which is used in Diff-A-Riff [7], M2L2 supports stereo inputs and outputs while achieving superior reconstruction quality at double the compression ratio.

3.2 Diffusion Transformers (DiT) replace the U-Net backbone used in image and audio-based diffusion models with a transformer-based architecture. DiT works by dividing the input representation into patches and processing them sequentially through transformer blocks. DiT uses Adaptive Layer Normalization (AdaLN), calculating normalization parameters from embeddings of diffusion timesteps and class labels. The AdaLN-Zero initialization strategy is shown to improve training stability and results further. In this work, we train a Latent DiT [40], [41] on sequences of latent embeddings from Music2Latent2. Latent DiTs were recently shown to generate long-form music with high resolution [5].

3.3 Consistency Models (CMs) [42], [43] are a new type of generative model capable of producing high-quality samples in a single forward pass, bypassing the need for adversarial training or iterative sampling. They learn a mapping from noisy to clean data via a probability flow ODE [44]. Given noise level σ, the consistency function f transforms a noisy sample $x_\sigma \sim p_\sigma(x)$ into a clean sample $x \sim p_{data}(x)$ using $f(x_\sigma, \sigma) \mapsto x$, typically parameterized as a neural network $f_\theta(x_\sigma, \sigma)$. To ensure $f_\theta(x, \sigma_{min}) = x$, where σ_{min} is the minimum noise level, CMs are expressed as:

$$f_\theta(x_\sigma, \sigma) = c_{skip}(\sigma)x_\sigma + c_{out}(\sigma)F_\theta(x_\sigma, \sigma),$$

with F_θ as a neural network, and $c_{skip}(\sigma)$, $c_{out}(\sigma)$ as differentiable functions satisfying $c_{skip}(\sigma_{min}) = 1$ and $c_{out}(\sigma_{min}) = 0$. Training involves discretizing the ODE over noise lev-

[2]Paper is under review. We provide an anonymized version at this link.

els $\sigma_1, \ldots, \sigma_N$ and minimizing the consistency loss:

$$\mathcal{L}_{CM} = \mathbb{E}[\lambda(\sigma_i, \sigma_{i+1}) d(f_\theta(x_{\sigma_{i+1}}, \sigma_{i+1}), f_{\theta^-}(x_{\sigma_i}, \sigma_i))],$$

where $d(x, y)$ is a distance metric, $\lambda(\sigma_i, \sigma_{i+1})$ scales the loss, and f_{θ^-} is a teacher network. After training, a sample x is generated in one step from noise $z \sim \mathcal{N}(0, I)$ using $x = f_\theta(z, \sigma_{max})$.

4 Method

4.1 Dataset. We use a proprietary dataset of over 20,000 multi-track recordings across various music genres and instruments, with a sample rate of 48 kHz. During training, we select a target accompaniment track (excluding vocals) and mix a random subset of remaining tracks to create the music context (*ctx*), following the same approach as in Diff-A-Riff [7]. We then divide training samples into 10-second windows. This process yields 1 million training pairs. A validation set is derived similarly from 2,000 recordings.

4.2 Experiments We conduct an ablation study to evaluate the impact of the modifications on Diff-A-Riff's original architectural and training setting. The experiments focus on the following enhancements:

- **Autoencoder Improvement (*M2L2*)**: We replace the original Music2Latent autoencoder used in Diff-A-Riff with Music2Latent2 [9], a state-of-the-art audio autoencoder featuring a transformer-based architecture and an autoregressive decoding scheme. Music2Latent2 supports stereo input and output, providing higher reconstruction quality at a 128× compression ratio—double that of the original Music2Latent—while maintaining the same latent dimensionality.
- **Architecture Transition (DiT)**: We shift from a convolutional U-Net to a Diffusion Transformer (DiT) [12] with depth-wise convolutions after Q, K, V projections in self-attention[45]. The DiT is initialized with a dimension of 1024, MLP multiplier of 4, 4 heads, and 18 layers. We use 512-dimensional sinusoidal embeddings [46] for noise levels, fed into AdaLN layers [12]. The model has around 280 million parameters (150 million less than Diff-A-Riff).
- **Consistency Training for Faster Inference (*C-DiT*)**: To enhance inference speed, we introduce consistency training to the latent model, following the framework in [43]. We use continuous noise levels and an exponential schedule for the consistency step, as shown in [9]. The remaining parameters are unchanged from the ones used for EDM training. Consistency training allows the model to generate high-quality audio with a few inference steps, significantly reducing computational overhead. We use 5 inference steps, which already improves quality over Diff-A-Riff (see Sec.5).
- **Bridging the Modality Gap (*CLAP*$_\beta$)**: To enhance the model's responsiveness to text prompts, we address the modality gap between audio-derived and text-derived CLAP embeddings [10], [13], [47], [48]. This gap seems to arise primarily due to the contrastive objective [49] and produces text and audio embeddings to lie in disjoint manifolds. To enable flexible sampling of multiple audio variations for a single text prompt—and to avoid the costly retraining of CLAP with new labels—we propose training a diffusion-based MLP model to reduce this gap. Leveraging a set of unstructured, human-annotated tags paired with our multi-track dataset (see Sec. 4.1), we train the MLP to predict audio-derived from text-derived CLAP embeddings, following a similar approach to [50]. The MLP is initialized with 1024 hidden units and uses eight residual dense blocks with Adaptive Layer Normalization (AdaLN) layers to integrate conditional text embeddings from CLAP. This setup effectively bridges the modality gap, improving generation quality for both text-prompted and unconditional audio CLAP embedding generation while allowing flexible sampling without requiring modifications to the CLAP model itself.

4.3 Implementation Details We train all models for 1 million iterations using AdamW [51], with $lr = 1e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We use a linear warmup during the first 1,000 steps, followed by a cosine decay until $lr = 0$ is reached. U-Net-based models are trained with a batch size of 256, while DiT-based models use a batch size of 128 for diffusion training and 16 for consistency training. Other details follow the same methodology as in Diff-A-Riff. We address a key issue of information leakage in Diff-A-Riff, where the CLAP embedding is derived from the same audio segment as the target. Instead, we extract the CLAP embedding from a random segment of the same target.

4.4 Evaluation We perform objective evaluations to assess the modifications made to Diff-A-Riff's original configuration [7]. Several metrics are used in this evaluation: *Kernel Distance* (KD) [52] and *Fréchet Audio Distance* (FAD) [53] assess audio quality while *Density* and *Coverage* [54] measure

fidelity and diversity. The Accompaniment Prompt Adherence (APA) [55] metric evaluates how well the generated accompaniment aligns with the given context.

The Clap Score (CS) [56] typically quantifies cross-modality similarities between individual pairs of text and audio in the CLAP space [10]. We refer to this metric as CS_{AT}. Additionally, we also report the *intra*-modality (CS_{AA}). In both cases, CS is computed based on generated data projected back into the CLAP space and the respective ground truth *embedding*. For $\textit{CLAP}_A$ conditioning, the ground truth for CS_{AA} is the reference audio, and the ground truth for CS_{AT} is the caption of the reference audio (which is not used for conditioning in the $\textit{CLAP}_A$ case). For $\textit{CLAP}_T$ conditioning, the ground truth for CS_{AA} is the audio whose caption was used for $\textit{CLAP}_T$ conditioning, and the ground truth for CS_{AT} is the actual caption used for computing the $\textit{CLAP}_T$ conditioning. For context-only conditioning (*ctx*), the ground truth for CS_{AA} is the audio stem that was originally part of the context but was removed and for CS_{AT}, it is the caption of that removed stem.

All metrics are calculated by averaging five batches of 1000 candidate samples. We use CLAP [10] as the embedding space for metrics that compare distributions (like KD and FAD) using a reference set of 5,000 real audio examples.

5 Results

Results are summarized in Table 1, where we compare the performance of our enhanced models to the original Diff-A-Riff under various conditioning settings. Music examples are available here.

Note that the values for the original Diff-A-Riff are different than in the original publication (cf. [7]). This is because here, we compare Diff-A-Riff generations with original audio data (e.g., to compute FAD, Coverage, Density, etc.), while in the original publication, all comparisons were done between generations and original data after a music2latent roundtrip (i.e., original data was encoded and decoded with music2latent). Also note that in the original Diff-A-Riff paper, the text conditioning was done using prompts generated by ChatGPT, while in this paper, we only use human-annotated lists of tags. Besides that, we use the same model architecture and weights as in the original publication.

Overall, the results improve when successively adding M2L2, DiT and $\textit{CLAP}_\beta$ ($\textit{CLAP}_A{}^*$ is a special case, where the unconditioned Modality-Gap Bridge diffusion model is used to generate CLAP audio embeddings). The results for *C-DiT* (i.e., the five-step latent consistency model) are slightly worse than with the conventional diffusion model but still substantially better than those of the original Diff-A-Riff model. We observe a slight decrease, though, in metrics such as APA and Clap Score (CS_{TA}). Consistency models are designed to generate samples in fewer steps by learning to approximate the denoising process more directly. However, this acceleration can come at the cost of reduced conditioning fidelity. One contributing factor is that consistency models cannot benefit as much from classifier-free guidance as traditional diffusion models. This limitation may lead to slightly lower alignment with the provided context and text prompts.

An interesting observation is that the best results in terms of KD^a, FAD, Density, and Coverage are achieved by $\textit{CLAP}_\beta$ using text conditioning ($\textit{CLAP}_T$). We hypothesize that this improvement comes from generating audio CLAP embeddings from text CLAP embeddings via our diffusion model $\textit{CLAP}_\beta$, which allows us to sample from a more diverse audio embedding distribution space and effectively transforms (unseen) text embeddings into audio embeddings that are within the conditional distribution the generative model was trained on. Consequently, this leads to improved performance on metrics that assess the quality and coverage of the generated samples.

The APA metric [55] reflects how well the generated audio aligns with the given audio context. The *+DiT* models achieve the best results in APA, demonstrating that these models are highly responsive to the provided audio context in generating accompaniments. Also, the new variants overall have a much better APA than the original Diff-A-Riff when only context conditioning is provided (*ctx*). This shows that the efficacy of the architecture improved considerably and that additional cues are not needed to produce valid accompaniments. This independence of conditionings is also striking when considering the improvements for fully unconditional generation (*uncond*).

When considering text-derived CLAP embeddings ($\textit{CLAP}_T$), the Clap Score (CS_{TA}) measures the alignment between the text prompts and the generated audio. The original Diff-A-Riff attains a CS_{TA} of 0.23 with both $\textit{CLAP}_T$ and Context, and 0.24 with $\textit{CLAP}_T$ only. Our enhanced models show a slight decrease in CS under the same settings. A possible reason for that is the fact that in the original Diff-A-Riff training, the CLAP audio embedding was computed from the exact target

	Inputs	↓ KD[a]	↓ FAD	↑ *Cov.*[b]	↑ *Den.*[b]	↑ APA	↑ CS_{AA}	↑ CS_{TA}
real	Original acc.	0.00	0.01	0.14	0.79	0.99	-	0.39
↓ *bound*	-	6.94[c]	1.65[c]	0.00[c]	0.00[c]	0.14[d]	-	0.13[d]
Diff-A-Riff	$CLAP_A$, *ctx*	1.48	0.42	0.07	0.42	0.85	0.53	0.13
	$CLAP_T$, *ctx*	1.59	0.46	0.08	0.52	0.96	0.47	**0.29**
	ctx	2.19	0.63	0.16	2.80	0.33	0.25	0.03
	$CLAP_A$	1.55	0.44	0.07	0.46	-	0.51	0.17
	$CLAP_T$	1.55	0.45	0.07	0.46	-	0.46	**0.29**
	uncond	2.46	0.70	0.12	3.46	-	0.23	0.02
+M2L2	$CLAP_A$, *ctx*	1.02	0.35	0.52	4.59	0.89	0.52	0.15
	$CLAP_T$, *ctx*	1.10	0.36	0.48	5.89	0.96	0.47	0.19
	ctx	0.99	0.34	0.61	7.07	0.88	0.41	0.09
	$CLAP_A$	1.11	0.38	0.44	4.13	-	0.49	0.13
	$CLAP_T$	1.10	0.38	0.45	5.06	-	0.45	0.16
	uncond	1.03	0.36	0.50	6.00	-	0.39	0.08
+DiT	$CLAP_A$, *ctx*	0.90	0.33	0.73	5.94	**1.00**	**0.53**	0.18
	$CLAP_T$, *ctx*	0.96	0.34	0.57	7.24	**1.00**	0.47	0.19
	ctx	**0.88**	**0.31**	0.62	7.53	0.99	0.42	0.11
	$CLAP_A$	1.04	0.38	0.61	5.39	-	0.49	0.15
	$CLAP_T$	1.04	0.38	0.57	7.29	-	0.44	0.15
	uncond	1.00	0.35	0.65	6.75	-	0.39	0.08
+$CLAP_\beta$	$CLAP_A{}^*$, *ctx*	0.93	0.33	0.65	6.51	0.99	0.50	0.15
	$CLAP_T$, *ctx*	**0.88**	0.32	**0.92**	**11.1**	0.87	0.49	0.19
	$CLAP_A{}^*$	1.00	0.35	0.58	6.38	-	0.48	0.14
	$CLAP_T$	0.93	0.34	0.81	10.1	-	0.47	0.17
C-DiT	$CLAP_A$, *ctx*	1.11	0.38	0.33	2.42	0.72	0.47	0.15
	$CLAP_T$, *ctx*	1.13	0.39	0.27	3.15	0.73	0.43	0.15
	ctx	1.08	0.36	0.37	3.73	0.67	0.39	0.10
	$CLAP_A$	1.27	0.43	0.25	2.01	-	0.44	0.11
	$CLAP_T$	1.29	0.43	0.20	2.41	-	0.39	0.13
	uncond	1.23	0.41	0.33	2.83	-	0.37	0.07

[a] $\times 10^{-3}$; [b] $\times 10^{-2}$
[c] obtained from white noise
[d] obtained using a random accompaniment from the dataset

Table 1: Objective metrics obtained for each configuration under different conditional settings. We compare against high-performance bounds obtained from the *real* validation set and low-performance bounds obtained from either *noise* or randomly paired music contexts and accompaniments. Some cells are empty for APA in the case of context-free generation.

segment, while here, we compute the embedding from a random segment of the target stem. This means that the new model variants generalize better to the instrument type and are less sensitive to the exact position of the CLAP embedding. Due to the gap between text and audio embeddings in CLAP, the generated audio is, therefore, less likely to project back exactly to the text embedding.

6 Conclusion

This work presents a series of substantial enhancements to Diff-A-Riff, a state-of-the-art latent diffusion model for musical accompaniment co-creation. By integrating a higher-fidelity stereo autoencoder, a transformer-based diffusion architecture, an advanced diffusion framework, and consistency training, we achieve significant improvements in audio quality, diversity, and inference speed. Additionally, our novel method for bridging the modality gap in CLAP embeddings enhances the model's responsiveness to text prompts, paving the way for more intuitive and precise text-driven audio generation. Through rigorous objective evaluation and ablation studies, we demonstrate the effectiveness of our proposed enhancements, highlighting the potential of our model for practical applications in music production. This work represents a further step towards developing AI-assisted tools that empower musicians with enhanced creative control and facilitate seamless integration of machine-generated content into their artistic workflows.

References

[1] A. Agostinelli *et al.*, "MusicLM: Generating Music From Text," in *CoRR*, 2023.

[2] J. Copet *et al.*, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems 36 NeurIPS*, 2023.

[3] Z. Evans *et al.*, "Fast timing-conditioned latent audio diffusion," *arXiv preprint arXiv:2402.04825*, 2024.

[4] Z. Evans *et al.*, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.

[5] Z. Evans *et al.*, "Stable audio open," *arXiv preprint arXiv:2407.14358*, 2024.

[6] H. Liu *et al.*, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," in *arXiv*, 2023.

[7] J. Nistal *et al.*, "Diff-a-riff: Musical accompaniment co-creation via latent diffusion models," *Proc. of the International Society for Music Information Retrieval (ISMIR)*, 2024.

[8] R. Rombach, A. Blattmann, *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[9] M. Pasini *et al.*, "Music2latent: Consistency autoencoders for latent audio compression," *arXiv preprint arXiv:2408.06500*, 2024.

[10] Y. Wu, K. Chen, *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, 2023.

[11] M. Pasini *et al.*, "Music2latent: Consistency autoencoders for latent audio compression," in *Proc. of the International Society for Music Information Retrieval (ISMIR)*, 2024.

[12] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023.

[13] W. Liang *et al.*, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Annual Conference on Neural Information Processing Systems NeurIPS*, 2022.

[14] A. van den Oord, S. Dieleman, *et al.*, "WaveNet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, Sep. 2016.

[15] S. Mehri, K. Kumar, *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," in *5th International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[16] I. J. Goodfellow, J. Pouget-Abadie, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Montreal, Quebec, Canada, Dec. 2014, pp. 2672–2680.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, Apr. 2014.

[18] J. Nistal *et al.*, "DrumGAN VST: A plugin for drum sound analysis/synthesis with autoencoding generative adversarial networks," in *Proc. of International Conference on Machine Learning ICML, Workshop on Machine Learning for Audio Synthesis, MLAS*, 2022.

[19] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," in *CoRR*, 2021.

[20] S. Rouard and G. Hadjeres, "CRASH: raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," in *Proc. of the 22nd International Society for Music Information Retrieval Conference, ISMIR*, 2021.

[21] A. van den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, Dec. 2017.

[22] P. Dhariwal, H. Jun, *et al.*, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[23] Y. Bai *et al.*, "Seed-music: A unified framework for high quality and controlled music generation," *arXiv preprint arXiv:2409.09214*, 2024.

[24] Y. Wu *et al.*, "Jukedrummer: Conditional beat-aware audio-domain drum accompaniment generation via transformer VQ-VAE," in *Proc. of the 3rd International Society for Music Information Retrieval Conference, ISMIR*, 2022.

[25] S.-L. Wu *et al.*, "Music controlnet: Multiple time-varying controls for music generation," in *CoRR*, 2023.

[26] Y. Zhang *et al.*, "Musicmagus: Zero-shot text-to-music editing via diffusion models," *CoRR*, 2024.

[27] H. Manor and T. Michaeli, "Zero-shot unsupervised and text-based audio editing using DDPM inversion," *CoRR*, 2024.

[28] Z. Novack *et al.*, "DITTO: diffusion inference-time t-optimization for music generation," *CoRR*, 2024.

[29] Z. Novack *et al.*, "Ditto-2: Distilled diffusion inference-time t-optimization for music generation," *arXiv preprint arXiv:2405.20289*, 2024.

[30] M. Levy *et al.*, "Controllable music production with diffusion models and guidance gradients," *CoRR*, 2023.

[31] M. Pasini *et al.*, "Bass accompaniment generation via latent diffusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, IEEE, 2024.

[32] C. Donahue *et al.*, "Singsong: Generating musical accompaniments from singing," in *CoRR*, 2023.

[33] J. D. Parker *et al.*, "StemGen: A music generation model that listens," in *CoRR*, 2023.

[34] S. Lattner and M. Grachten, "High-level control of drum track generation using learned patterns of rhythmic interaction," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, IEEE, 2019.

[35] M. Grachten *et al.*, "BassNet: A variational gated autoencoder for conditional generation of bass guitar tracks with learned interactive control," in *Applied Sciences*, 2020.

[36] E. Postolache *et al.*, "Generalized multi-source inference for text conditioned music diffusion models," *CoRR*, 2024.

[37] G. Mariani *et al.*, "Multi-source diffusion models for simultaneous music generation and separation," *CoRR*,

[38] P. E. Giovanni Bindi, "Unsupervised composable representations for audio," *Proc. of the International Society for Music Information Retrieval (ISMIR)*, 2024.

[39] T. Karchkhadze *et al.*, "Simultaneous music separation and generation using multi-track latent diffusion models," *CoRR*, vol. abs/2409.12346, 2024.

[40] P. Esser, S. Kulal, *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," *arXiv preprint arXiv:2403.03206*, 2024.

[41] M. Haji-Ali *et al.*, "Taming data and transformers for audio generation," *arXiv preprint arXiv:2406.19388*, 2024.

[42] Y. Song *et al.*, "Consistency models," in *International Conference on Machine Learning, ICML*, 2023.

[43] Y. Song and P. Dhariwal, "Improved techniques for training consistency models," *arXiv preprint arXiv:2310.14189*, 2023.

[44] J. Song *et al.*, "Denoising diffusion implicit models," in *Proc. of the 9th International Conference on Learning Representations, ICLR*, 2021.

[45] D. R. So *et al.*, "Searching for efficient transformers for language modeling," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato *et al.*, Eds., 2021, pp. 6010–6022.

[46] A. Vaswani, N. Shazeer, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, Dec. 2017.

[47] H. Wu *et al.*, "Audio-text models do not yet leverage natural language," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, IEEE, 2023, pp. 1–5.

[48] S. Ghosh *et al.*, "Compa: Addressing the gap in compositional reasoning in audio-language models," *arXiv preprint arXiv:2310.08753*, 2023.

[49] A. Fahim *et al.*, "Its not a modality gap: Characterizing and addressing the contrastive gap," *CoRR*, vol. abs/2405.18570, 2024.

[50] A. Ramesh, P. Dhariwal, *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, 2022.

[51] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR*, 2019.

[52] M. Binkowski *et al.*, "Demystifying MMD gans," in *Proc. of the 6th International Conference on Learning Representations, ICLR*, 2018.

[53] K. Kilgour, M. Zuluaga, *et al.*, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, Sep. 2019.

[54] M. F. Naeem *et al.*, "Reliable fidelity and diversity metrics for generative models," in *Proc. of the 37th International Conference on Machine Learning, ICML*, 2020.

[55] M. Grachten and J. Nistal, "Audio Prompt Adherence: A measure for evaluating musical accompaniment systems," in *CoRR*, 2024.

[56] R. Huang *et al.*, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. of the International Conference on Machine Learning, ICML*, 2023.