

# A Training-Free Approach for Music Style Transfer with Latent Diffusion Models

Sooyoung Kim<sup>1\*</sup>, Joonwoo Kwon<sup>1\*</sup>, Heehwan Wang<sup>1\*</sup>,  
Shinjae Yoo<sup>2†</sup>, Yuewei Lin<sup>2†</sup>, Jiook Cha<sup>1†</sup>

<sup>1</sup>Seoul National University, Republic of Korea    <sup>2</sup>Brookhaven National Laboratory, Upton, NY, USA  
{rlatndud0513, pioneers, dhkdgmlghks}@snu.ac.kr,  
{sjyoo, ywlin}@bnl.gov, connectome@snu.ac.kr,

## Abstract

Music style transfer, while offering exciting possibilities for personalized music generation, often requires extensive training or detailed textual descriptions. This paper introduces a novel training-free approach leveraging pre-trained Latent Diffusion Models (LDMs). By manipulating the self-attention features of the LDM, we effectively transfer the style of reference music onto content music without additional training. Our method achieves superior style transfer and melody preservation compared to existing methods. This work opens new creative avenues for personalized music generation.

## Introduction

Music has long been a fundamental element of human culture, offering unique experiences that resonate with individual listeners (Barton 2018). As a universal language, music transcends boundaries by facilitating both communication of ideas (Miell, MacDonald, and Hargreaves 2005) and emotional expression through various forms of engagement, from composition to performance and listening (Robinson and Hatten 2012). Nonetheless, music creation has traditionally presented significant barriers to entry, requiring specialized knowledge of theory, instrumental mastery, and technical expertise in composition and production. These requirements have historically restricted music production to individuals with formal training or substantial resources.

The rise of artificial intelligence, with its goal of emulating human creativity, has sparked a significant increase in interest in music-related research. Early efforts focused on challenges such as representing music data (Wiggins 1995; Camurri et al. 1995; Balaban 1996) and generating individual notes (sounds) (Miranda 1995). Over time, the scope of research expanded to include more complex issues, such as the generation of longer musical segments, music classification (Weihs et al. 2007; Fernández and Vico 2013; Kaliakatsos-Papakostas, Floros, and Vrahatis 2020; Ndou, Ajoodha, and Jadhav 2021), and music recommendation (Casey et al. 2008; Song, Dixon, and Pearce 2012). Building upon these advancements, recent developments in AI have led to the exploration of more sophisticated techniques,

such as music style transfer. This innovation has opened up new possibilities for ordinary individuals, enabling them to create and experience personalized music in ways that were previously inaccessible.

Inspired by the success of neural style transfer in computer vision (Gatys 2015; Johnson, Alahi, and Fei-Fei 2016; Chandran et al. 2021; Kwon et al. 2024; Wang et al. 2023), music style transfer has emerged as a powerful technique for generating music by decomposing and recombining the content and style elements of different musical pieces (Dai, Zhang, and Xia 2018). Though the distinction between musical content and musical style is not formally defined and context-dependent, researchers have proposed various frameworks for decomposing musical elements into content and style components. Content typically encompasses structural elements such as melody, harmony, and rhythmic patterns, while style encompasses performance-specific attributes including timbre, articulation, dynamics, and genre-specific characteristics (Dai, Zhang, and Xia 2018; Cífka, Şimşekli, and Richard 2019). Several studies further demonstrated the possibility of style transfer using non-musical elements, such as environmental sound (Grinstein et al. 2018).

Recent advances in large language models (LLMs) and diffusion models have simplified music generation and style transfer, enabling intuitive control even for those without detailed musical knowledge. Conventional methods for diffusion model-based music generation utilize the generative power of pre-trained diffusion models. Agostinelli et al. (2023) and Copet et al. (2024) introduced music generation using text conditioning (text-to-audio generation) and melody guidance by using a transformer-based autoregressive model. These works make style transfer more simple and controllable by using text descriptions as style specifications and audio clips as content. Most recently, drawing inspiration from textual inversion in the diffusion model (Gal et al. 2022), Li et al. (2024) utilized a pseudo-word representation for musical styles, enabling style transfer for arbitrary input music while maintaining structural information (e.g., melody or rhythm). However, to obtain satisfactory style-transfer results, it is necessary to provide detailed textual guidance on various musical attributes (e.g., timbre, pitch, performance style, composition style), which require specialized knowledge. Additionally, training or fine-tuning these models demands substantial computational power and

\*These authors contributed equally.

†Co-corresponding authors.

time, which greatly limits their practical application in real-world scenarios.

To overcome these limitations, we propose a straightforward yet effective training-free approach for music style transfer using pre-trained latent diffusion models (LDM) (Rombach et al. 2022), which are commonly used for training text-to-image diffusion models. Building on the findings of (Chung, Hyun, and Heo 2024; Feng et al. 2022; Ma et al. 2024), attention maps are responsible for determining the spatial arrangement, while the *key* and *value* in cross-attention handle the content that fills the space. Additionally, (Chung, Hyun, and Heo 2024) showed that the self-attention layer in diffusion models is well-suited for style transfer, as it preserves the relationships between content image patches after the transfer and encourages style transfer based on the similarity of local textures between the content and the style. As a result, our method aims to transfer the style of reference music to the content music by explicitly manipulating the self-attention features of pre-trained large-scale diffusion models for text-to-image synthesis without any further training or optimization while leveraging the image characteristics of mel-spectrograms. Specifically, we simply replace the content’s *key* and *value* of self-attention with those of the style music, focusing particularly on the later layers of the decoder that capture relevant local textures of mel-spectrograms. To further improve music stylization, we also incorporate additional techniques originally introduced in (Chung, Hyun, and Heo 2024), including query preservation, attention temperature scaling, and initial latent Adaptive Instance Normalization (AdaIN).

Our key contributions include:

- We propose a straightforward yet effective training-free approach for music style transfer, leveraging pre-trained LDM through direct manipulation of self-attention features.
- Extensive experiments validate that our method can faithfully transfer the style of reference music to content music without the need for additional training or fine-tuning.

## Related Works

**Music Style Transfer.** Music style transfer enables musical generation through element decomposition and recombination (Dai, Zhang, and Xia 2018). The field has evolved through various neural architectures, each addressing different aspects of musical transformation. Initial research focused on timbre transformation while preserving melodic content, with Engel et al. (2017) using WaveNet-based autoencoders and Grinstein et al. (2018) applying Convolutional Neural Networks (CNN) for timbre transfer between musical and non-musical sounds. These successes extended to genre transformation, where Brunner et al. (2018) demonstrated symbolic music manipulation using Recurrent Neural Networks (RNN). Cífka, Şimşekli, and Richard (2019) achieved more sophisticated genre transformation by separately processing melodic content and arrangement style by using the hybrid of CNN and RNN. Adversarial training approaches also produced significant results. WaveNet and CycleGAN architectures enabled style-preserved timbre trans-

fer (Huang et al. 2018; Bonnici, Benning, and Saitis 2022), while Hung et al. (2019) achieved direct disentanglement of pitch and timbre. By leveraging adversarial approach, (Lee et al. 2020) demonstrated the potential of multi-media style transfer by injecting music as the style information for image style transfer.

Music style transfer has advanced significantly through both LLM and diffusion-based approaches. LLM-based methods, such as MusicLM (Agostinelli et al. 2023) and MusicGen (Copet et al. 2024), introduced transformer-based autoregressive models that enable text-conditioned and melody-guided generation, making style control more accessible through text descriptions and audio references. Huang et al. (2023) (Huang et al. 2023) first demonstrated diffusion models’ potential in music generation, working with both waveforms and spectrograms. Building on these advances and drawing inspiration from textual inversion image diffusion models (Gal et al. 2022), Li et al. (2024) (Li et al. 2024) developed a pseudo-word representation approach, achieving structure-preserving style transfer that maintains melodic and rhythmic elements with flexible style manipulation for arbitrary input music. However, all of these algorithms require detailed guidance or extensive model training. In this paper, we faithfully transfer the style to the content music without any optimization process.

**Diffusion Models.** Diffusion models (DM) are a subclass of generative models based on likelihood estimation, with the foundational work being the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020). These models are grounded in the theoretical principles of Markov chains and Langevin dynamics. Due to their stable training process and scalability, diffusion models have surpassed Generative Adversarial Networks (GANs) (Dhariwal and Nichol 2021) in image generation tasks, yielding superior sample quality. However, the sampling process in diffusion models is typically slow, as it necessitates the generation of samples through a stepwise Markov chain process. To mitigate this issue, Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2020) introduce a non-Markovian iterative sampling method that accelerates the process while preserving the training procedure. More recently, the LDM (Rombach et al. 2022) has been proposed for image synthesis. This approach compresses images into a lower-dimensional latent space before applying the diffusion process, significantly reducing computational complexity while maintaining high-quality image generation. However, the use of LDM for music generation remains a relatively underexplored area, primarily due to the challenges posed by the scarcity of relevant data and the significant computational cost associated with model training. To address these limitations, we propose a novel, training-free method for music style transfer that does not require any specific dataset or additional training. To the best of our knowledge, this is the first attempt to apply an attention-based manipulation method for style infusing within the context of diffusion models for music style transfer.

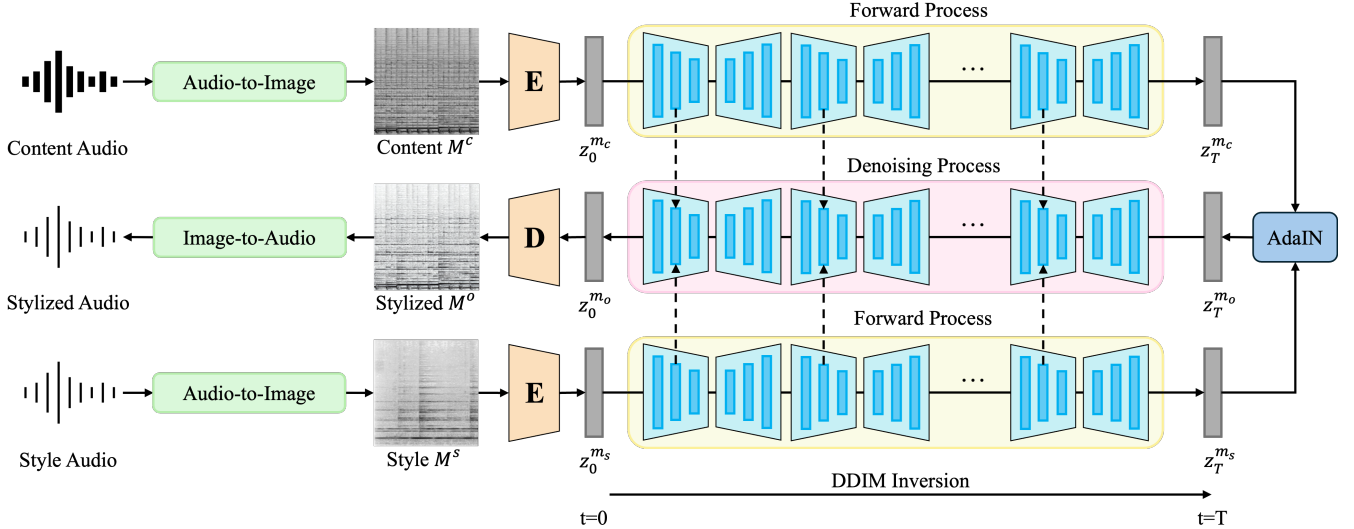


Figure 1: **Overall architecture** for our approach. Note that our approach does not require a training process

## Method

We utilized Stable Diffusion (Rombach et al. 2022) as the backbone to achieve music stylization, as shown in Figure 1. Our work is conducted in the image domain, processing a mel-spectrogram obtained from the input audio waveform using a Short Time Fourier Transform (STFT), based on the ideas of Riffusion (Forsgren and Martiros 2022) and MusicTI (Li et al. 2024).

**The Self-Attention Block in LDM.** The Latent Diffusion Model (LDM) (Rombach et al. 2022) is a type of diffusion model trained in a lower-dimensional latent space, which enables the model to focus on essential semantic features of data while reducing computational complexity. Given an image  $x \in \mathbb{R}^{H \times W \times 3}$ , the encoder  $\mathcal{E}$  encodes  $x$  into the latent representation,  $z \in \mathbb{R}^{h \times w \times c}$ , and the decoder reconstructs the image from the latent space.

Leveraging a pre-trained encoder, the entire images are encoded into latent space, and a diffusion model is trained on these latent representations,  $z$ . The model predicts the noise  $\epsilon$  added to the noised version of the latent variable  $z_t$  at each time step  $t$ . The training objective for LDM is given by:

$$L_{LDM} = \mathbb{E}_{z, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2], \quad (1)$$

where  $\epsilon \in \mathcal{N}(0, 1)$  represents noise,  $t$  is a time step uniformly sampled from  $\{1, \dots, T\}$ , and  $y$  is a condition.  $\epsilon_\theta$  is a neural network that predicts the noise added to  $z$ . In our case,  $y$  is a text condition, and  $\epsilon_\theta$  is modeled using a U-Net architecture that includes a residual block, a self-attention block, and a cross-attention block for each resolution in sequence.

Given a feature  $\phi$  after the residual block and a projection layer  $f(\cdot)$ , the self-attention mechanism is computed as follows:

$$Q = f_q(\phi), K = f_k(\phi), V = f_v(\phi),$$

$$\phi_{out} = \text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \quad (2)$$

where  $d$  represents the feature dimension of the projected *query*.  $Q, K, V$  represent *query, key, and value*, respectively. Notably, in our configuration, the condition  $y$  is an empty text prompt, meaning that no specific text conditions are applied.

## Attention-based Style Manipulation

Following (Chung, Hyun, and Heo 2024), we treat the features in the self-attention layers similarly to cross-attention, using the style image  $I^s$  as the conditioning input. Specifically, during the generation process, we replace the *key* and *value* features of the content music’s mel-spectrogram with those from the style music’s mel-spectrogram. To achieve this, we first obtain the latent representations for both the content and style mel-spectrograms through DDIM inversion (Song, Meng, and Ermon 2020). We then capture the self-attention features of the style mel-spectrogram throughout the DDIM inversion process. For pre-defined timesteps  $t = 0, \dots, T$ , we invert the style and content mel-spectrograms, denoted as  $z_0^{m_c}$  and  $z_0^{m_s}$ , from the image space ( $t = 0$ ) to Gaussian noise ( $t = T$ ). During this process, we also gather the *query* features of the content mel-spectrogram ( $Q_t^{m_c}$ ) and the *key* and *value* features of the style mel-spectrogram ( $K_t^{m_s}, V_t^{m_s}$ ) at each timestep. Next, we initialize the stylized output latent noise  $z_T^{m_o}$  by directly copying the content latent noise  $z_T^{m_c}$ . The style transfer is carried out by replacing the original *key*  $K_t^{m_o}$  and *value*  $V_t^{m_o}$  in the self-attention layer with the *key*  $K_t^{m_s}$  and *value*  $V_t^{m_s}$  derived from the style mel-spectrogram, during the reverse process of generating the stylized output latent  $z_T^{m_o}$ . To preserve the content

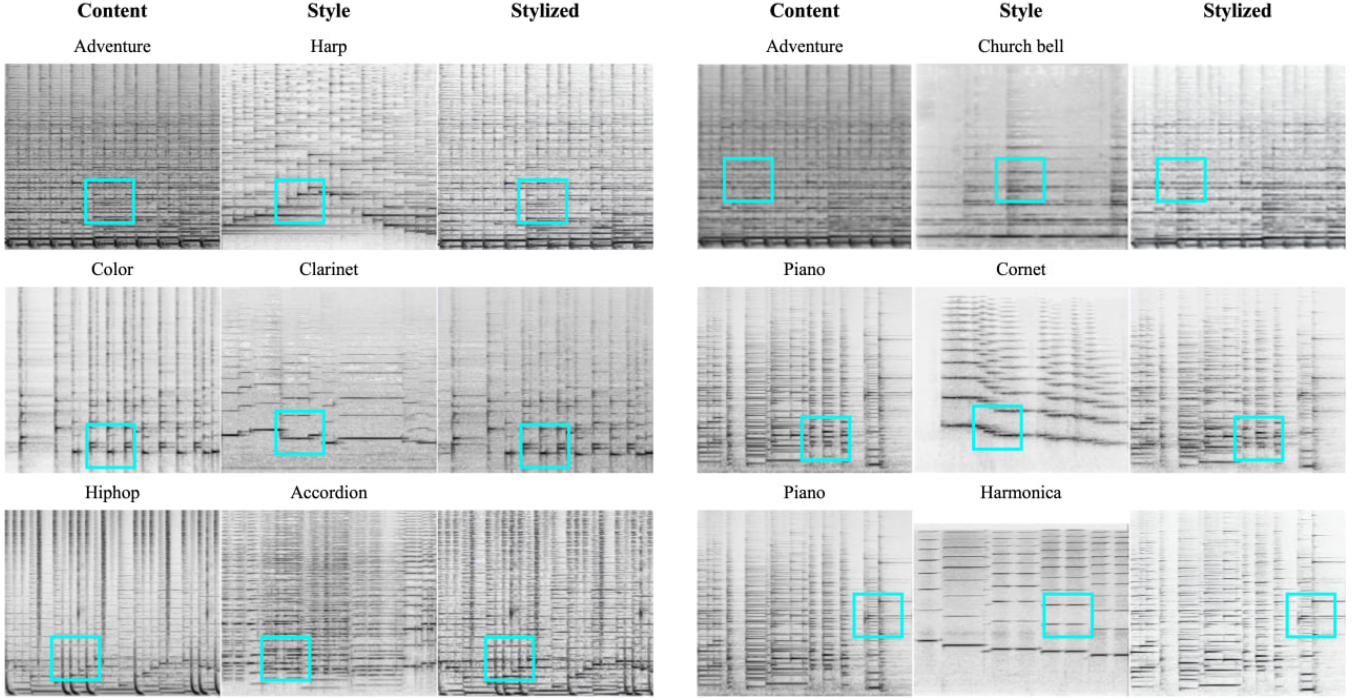


Figure 2: Qualitative comparisons through mel-spectrogram visualizations. The plots clearly demonstrate that the target domain has been effectively shifted while the content stays preserved.

structure and avoid unwanted distortions, we implement a query preservation technique as follows:

$$\begin{aligned}\bar{Q}_t^{m_o} &= \alpha \times Q_t^{m_c} + (1 - \alpha) \times Q_t^{m_o}, \\ \phi_{out}^{m_o} &= \text{Attn}(\bar{Q}_t^{m_o}, K_t^{m_s}, V_t^{m_s}),\end{aligned}\quad (3)$$

where  $\alpha$  is a hyperparameter. These operations are applied to the later layers of the decoder (layers 7–12 in the SD model) that focus on capturing local texture features.

**Additional Techniques.** In line with the approach of (Chung, Hyun, and Heo 2024), we applied an attention temperature scaling to rectify the attention map sharper and employed AdaIN to modulate the initial latent for guiding the generation process to capture the structural features of a mel-spectrogram from the content, as expressed in:

$$z_T^{m_o} = \sigma(z_T^{m_s}) \left( \frac{z_T^{m_c} - \mu(z_T^{m_c})}{\sigma(z_T^{m_c})} \right) + \mu(z_T^{m_s}), \quad (4)$$

where  $\mu(\cdot), \sigma(\cdot)$  denote channel-wise mean and standard deviation, respectively. Our empirical results show that using AdaIN for modulating the initial latent provides optimal performance compared to other style transfer methods, such as AdaConv (Chandran et al. 2021) or EFDN (Zhang et al. 2022).

### Implementation Details

We experiment with the MusicTI Dataset (Li et al. 2024), which contains a total of 254 five-second clips, with 74 style clips and 179 content clips. Note that our approach does not require a training process; thus, we used all these data at

inference only. For LDM, we used stable diffusion ver.1.5 (Rombach et al. 2022). In all our experiments, we fix the parameters of LDM and use the author-released codes using default configurations. All experiments were conducted using the PyTorch framework (Paszke et al. 2019) on a single NVIDIA A100(40G) GPU.

## Experimental Results

In this section, the proposed model’s validity is assessed both qualitatively and quantitatively.

Figure 2 illustrates that our approach produces qualitatively satisfactory audio-transferred outputs throughout various instruments and musical elements. Analysis of mel-spectrograms shows that stylized outputs maintain content structure while incorporating characteristic elements of style sources. For instance, when applying accordion style to hip-hop audio, the stylized mel-spectrograms display distinctive horizontal energy bands in mid-frequencies, characteristic of accordion harmonics. Similarly, cornet-styled piano pieces exhibit concentrated mid-frequency energy and reduced high-frequency components, reflecting the cornet’s distinctive timbral characteristics. It is worth noting that even though the stable diffusion model is originally trained for text-to-image generation, not for music generation, it showed remarkable performance on synthesizing style-transferred music without any optimization process and music datasets.

	content	style
$FAD_{vgg} (\downarrow)$	9.46	24.49
$FAD_{CLAP} (\downarrow)$	1.10	1.23

Table 1: Quantitative results on FAD scores using pre-trained VGG and CLAP.

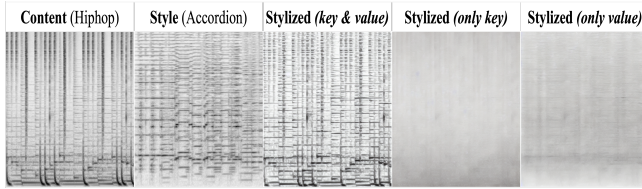


Figure 3: **The effectiveness of attention-based music-style manipulation.** Without the guidance from the style mel-spectrogram, the output shows a blurry mel-spectrogram, indicating unclear and unrefined results. This highlights the importance of the *key* and *value* from the style mel-spectrogram’s self-attention layers being crucial for generating high-quality outputs.

## Conclusion & Future Works

We present a novel approach to music style transfer that leverages pre-trained LDM through attention manipulation, enabling musical style transfer without additional training. Our method modifies self-attention layers during content generation by replacing content’s key and value components with those from style music, similar to cross-attention mechanisms. Our experimental results demonstrate successful style injection between mel-spectrograms without additional training. This extensibility of our approach suggests broad applications across various audio domains, including speech and environmental sounds, which we plan to explore in future work. Furthermore, we plan to comprehensively evaluate our approach with multiple evaluation metrics.

## References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Balaban, M. 1996. The music structures approach to knowledge representation for music processing. *Computer Music Journal*, 20(2): 96–111.

Barton, G. 2018. *Music learning and teaching in culturally and socially diverse contexts: Implications for classroom practice*. Springer.

Bonnici, R. S.; Benning, M.; and Saitis, C. 2022. Timbre transfer with variational auto encoding and cycle-consistent adversarial networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Brunner, G.; Konrad, A.; Wang, Y.; and Wattenhofer, R. 2018. MIDI-VAE: Modeling dynamics and instrumentation

of music with applications to style transfer. *arXiv preprint arXiv:1809.07600*.

Camurri, A.; Catorcini, A.; Innocenti, C.; and Massari, A. 1995. Music and multimedia knowledge representation and reasoning: the harp system. *Computer music journal*, 19(2): 34–58.

Casey, M. A.; Veltkamp, R.; Goto, M.; Leman, M.; Rhodes, C.; and Slaney, M. 2008. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4): 668–696.

Chandran, P.; Zoss, G.; Gotardo, P.; Gross, M.; and Bradley, D. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7972–7981.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

Cífka, O.; Şimşekli, U.; and Richard, G. 2019. Supervised symbolic music style translation using synthetic data. *arXiv preprint arXiv:1907.02265*.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Dai, S.; Zhang, Z.; and Xia, G. G. 2018. Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Engel, J.; Resnick, C.; Roberts, A.; Dieleman, S.; Norouzi, M.; Eck, D.; and Simonyan, K. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, 1068–1077. PMLR.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

Fernández, J. D.; and Vico, F. 2013. AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, 48: 513–582.

Forsgren, S.; and Martiros, H. 2022. Riffusion - Stable diffusion for real-time music generation.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gatys, L. A. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Grinstein, E.; Duong, N. Q.; Ozerov, A.; and Pérez, P. 2018. Audio style transfer. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 586–590. IEEE.

- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Huang, Q.; Park, D. S.; Wang, T.; Denk, T. I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; et al. 2023. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.
- Huang, S.; Li, Q.; Anil, C.; Bao, X.; Oore, S.; and Grosse, R. B. 2018. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620*.
- Hung, Y.-N.; Chiang, I.; Chen, Y.-A.; Yang, Y.-H.; et al. 2019. Musical composition style transfer via disentangled timbre representations. *arXiv preprint arXiv:1905.13567*.
- Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 694–711. Springer.
- Kaliakatsos-Papakostas, M.; Floros, A.; and Vrahatis, M. N. 2020. Artificial intelligence methods for music generation: a review and future perspectives. *Nature-Inspired Computation and Swarm Intelligence*, 217–245.
- Kwon, J.; Kim, S.; Lin, Y.; Yoo, S.; and Cha, J. 2024. AesFA: An Aesthetic Feature-Aware Arbitrary Neural Style Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13310–13319.
- Lee, C.-C.; Lin, W.-Y.; Shih, Y.-T.; Kuo, P.-Y.; and Su, L. 2020. Crossing you in style: Cross-modal style transfer from music to visual arts. In *Proceedings of the 28th ACM international conference on multimedia*, 3219–3227.
- Li, S.; Zhang, Y.; Tang, F.; Ma, C.; Dong, W.; and Xu, C. 2024. Music Style Transfer with Time-Varying Inversion of Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 547–555.
- Ma, W.-D. K.; Lahiri, A.; Lewis, J. P.; Leung, T.; and Kleijn, W. B. 2024. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4098–4106.
- Miell, D.; MacDonald, R. A.; and Hargreaves, D. J. 2005. *Musical communication*. Oxford University Press, USA.
- Miranda, E. R. 1995. An artificial intelligence approach to sound design. *Computer Music Journal*, 19(2): 59–75.
- Ndou, N.; Ajoodha, R.; and Jadhav, A. 2021. Music genre classification: A review of deep-learning and traditional machine-learning approaches. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 1–6. IEEE.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Robinson, J.; and Hatten, R. S. 2012. Emotions in music. *Music Theory Spectrum*, 34(2): 71–106.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Song, Y.; Dixon, S.; and Pearce, M. 2012. A survey of music recommendation systems and future perspectives. In *9th international symposium on computer music modeling and retrieval*, volume 4, 395–410. Citeseer.
- Wang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Chen, H.; Xing, W.; and Lu, D. 2023. MicroAST: Towards Super-Fast Ultra-Resolution Arbitrary Style Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Weihs, C.; Ligges, U.; Mörchen, F.; and Müllensiefen, D. 2007. Classification in music research. *Advances in Data Analysis and Classification*, 1: 255–291.
- Wiggins, G. A. 1995. Understanding music with AI—Perspectives on cognitive musicology.
- Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022. Exact Feature Distribution Matching for Arbitrary Style Transfer and Domain Generalization. In *CVPR*.