An Image-to-Music Generation Framework Powered by An Algorithm-Driven Music Core

Callie C. Liao
IntelliSky
McLean, VA USA
ccliao@intellisky.org

Duoduo Liao
George Mason University
Fairfax, VA USA
dliao2@gmu.edu

Ellie L. Zhang
IntelliSky
McLean, VA USA
elzhang@intellisky.org

Abstract

With the rise of generative AI, music generation, particularly image-to-music, remains underexplored and relies on deep learning methods. Existing neural-based models depend on large datasets, raising concerns about copyright infringement and increased costs for performance improvements. In contrast, we propose an innovative image-to-music AI framework powered by a novel, algorithm-driven music core, minimizing copyright infringement risks. Our music core connects lyrical and rhythmic information to automatically derive musical features, constructing a complete score from lyrics. In this pilot study, we developed a web tool based on this framework that generates melodies adhering to music theory, lyrical, and rhythmic conventions. Experimental results show that our approach achieves an average music key confidence score of 0.86, surpassing the 0.8 score of human composers and demonstrating its ability to produce diverse, human-like compositions. Therefore, this tool serves as a reliable co-pilot for composers as well as entertainment, advancing AI in music generation.

Click on the web link to run our web tool, GenAIM, to generate music from images.

1. Introduction

Generative AI has experienced rapid escalation and integration into daily lives, particularly with the frequent use of conversational chatbots such as ChatGPT [1] powered by Large-Language Models (LLMs). However, AI music generation, especially multimodal inputs from images, lags behind AI art and writing due to its complex structure and required musical expertise. Current methods rely on deep learning [2][4][11], using large datasets for music generation, but face challenges including data collection, copyright risks, high computing costs, and labor-intensive data preparation [3][7][9].

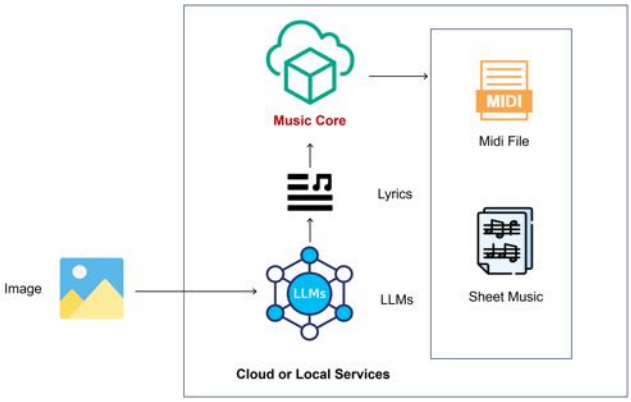

Figure 1. The System Architecture of the Framework.

To develop a musically robust generation method, background in music theory, literature, and linguistics is required to emulate the intuitive thinking process of musical artists to remove the necessity of relying on existing music for algorithm training and development. Therefore, we propose a novel image-to-music framework driven by an innovative pure-algorithm-powered music core. Our core method purely utilizes novel algorithms based on lyrical input, addressing the challenges listed above by leveraging lyric-music correlations, as inspired by [13][15]. Our method also discovers rhythmic, syllabic, and stress patterns to ensure proper lyric-music alignment. This framework allows users, regardless of their expertise, to generate music for reference or entertainment, with customizable features such as key signatures, instruments, and sheet music display. The ability to use images to generate music creates another input option that expands the degree of human and AI creativity when creating AI-generated music. Examples are shown in Figure 4.

This paper presents a non-neural approach to AI music generation from images, combining lyrics, melody, music theory, and composer intuition to create natural, human-sounding music. Our main contributions are as follows:

- Our framework generates melodies from image-derived text using novel, non-deep learning music core algorithms, differentiating it from existing models.
- Our approach does not require training data, avoiding copyright issues, minimizing manual labor, and ensuring cost-effectiveness.
- Through the understudied method of pairing keywords with strong beats, our core music generation achieves lyrical alignment with the rhythmic structure.
- Our image-to-music approach uses LLMs to generate lyrics, followed by purely algorithmic music composition.
- As a co-pilot tool, it empowers composers and aspiring musicians, offering benefits for entertainment, creativity, and overall well-being.

2. Related Work

Currently, image-to-music generation remains largely unexplored. Research in this area involves deep learning methods that extract essential visual attributes, which are subsequently translated into melodies, harmonies, and rhythmic patterns [6]. Emotion is heavily considered in various ways. Some utilize the valence-arousal emotional space to detect the emotional tone of an image, while other researchers perform image analysis assuming that the components are musically related [19][10][18]. These works emphasize the potential emotions that could be extracted from the images. Recent research methodologies [5][17][12] focus on generating audio music from multiple modalities, including images and videos. MelFusion [5] synthesizes music from image and language cues using diffusion models. Diff-BGM [12] also uses a diffusion model to generate background music for videos. However, our image-to-music methodology distinguishes from these by utilizing LLM technologies to generate lyrics before employing our purely algorithmic lyric-to-music approach, reducing the need to specifically analyze emotions from the image. Our image-to-music approach differs by utilizing LLMs to generate lyrics from an input image, followed by the use of our music core to generate the corresponding music.

3. The Framework

This framework is built on the AWS platform. The input image is processed by the LLMs provided by the AI services, which generate the lyrics. These generated lyrics are then passed to the music core for music composition. The front-end web application loads the music files from the AWS cloud, renders it to a music sheet, and plays the music. Figure 1 presents the system architecture of the proposed framework. The music core algorithms, our main focus, are explained in the following.

3.1. Music Core Algorithms

3.1.1. Score Setup

To set up the score, lyrical information, including syllables and keywords, is extracted to determine the time signature, as described in [14]. Phrases are identified through punctuation with accents and keywords influencing the total number of measures.

3.1.2. Rhythmic Score Construction

The time signature is essential for defining the rhythmic structure and distributing phrases across measures. Keywords are inserted into stressed beats within each measure following the order specified by the lyrics. This approach, based on the connection between keywords and stressed beats as identified in [13][15], is incorporated into the proposed framework.

3.1.3. Pitch Construction

During the pitch construction, key signature, pitch range, and phrase length are first defined: the key sets the tonal center and mode (e.g., D major), either chosen by the user or selected randomly based on the highest key confidence. The pitch range, limited to a certain singing range, ensures a comfortable singing range, while the phrase length guides melodic phrasing. The pitch generation and insertion processes work in a feedback loop, with randomly generated pitches adapted to music theory before being refined for smoother transitions and varied phrasing to avoid repetition and enhance melody flow.

3.1.4. XML Score Conversion

After constructing the score and pitch, the music is converted into a MusicXML file for better readability and compatibility with composition and notation software, facilitating digital sheet music exchange and collaboration.

4. Experimental Results & Evaluation

In this pilot study, we developed a web tool, GenAIM, based on our proposed framework that generates melodies adhering to music theory, lyrical, and rhythmic conventions. Then, we performed exploratory data analysis using Music21 [16], a Python-based development toolkit created by the Massachusetts Institute of Technology (MIT)[16][8], to identify similarities, differences, and trends between AI-generated and original compositions.

To evaluate the generated music in comparison to human-composed pieces, we used 25 original lyrics from piano books and generated 4 to 5 songs per lyrical song group in various keys, resulting in 112 AI-generated pieces. Each set included at least one version matching the time and key signature of the original song, enabling direct comparison for structural and musical analysis.

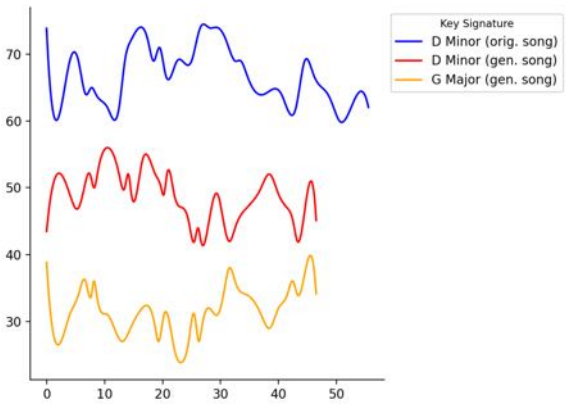

Figure 2. Melody motion of Generated Songs vs. Original Songs

4.1. Melodic Motion

Figure 2 presents motion plots of three AI-generated songs in different key signatures, with x-axis values representing relative note positions and y-axis values indicating pitch positions within each key. Vertical shifts were applied for clarity, so pitch trends should only be compared within the same key. The turning points along the x-axis closely align with those of the original songs, indicating similar rhythmic patterns, while greater pitch variation is expected due to the melodic flexibility allowed in music composition.

4.2. Key Confidence

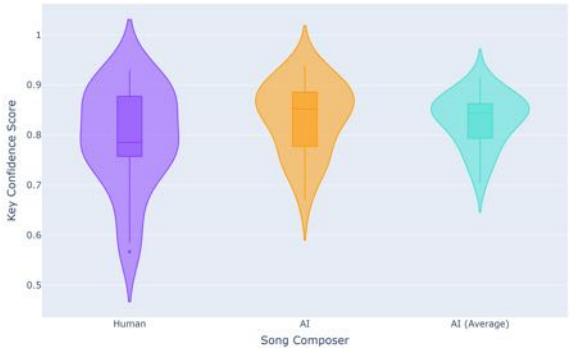

Figure 3. Key confidence of generated songs vs. original songs with various key signatures comparing.

Figure 3 shows a violin plot comparing key confidence scores—derived from music21 algorithms—across human composers, an AI composer, and an averaged AI composer. The AI composer retains the original key and time signatures, while the averaged AI composer generates songs without those constraints. Human compositions show a broader distribution, whereas AI composers produce more concentrated scores, particularly in the 0.8–0.9 range, suggesting greater consistency and alignment with a well-defined tonal center in AI-generated music.

4.3. Rhythm Matching

The final evaluation focuses on rhythm matching, which is essential for accurate syllable placement and emphasizing keywords through duration and beat position. The generated songs achieved an average rhythm matching accuracy of 73.6% compared to the original songs, indicating a relatively high level of alignment. Some deviation is expected due to the inherent flexibility in song composition; however, research from [13][15] recommends associating keywords with strong beats to maintain musical and lyrical coherence.

5. Conclusions & Future Work

This paper presents a state-of-the-art image-to-music generation framework that enables image-to-music composition without relying on prior music training data. Our framework supports the creation of both lyrical and instrumental pieces, and its effectiveness is demonstrated through a comparative analysis of 112 AI-generated songs against 25 original human-composed works. The results show that our framework produces melodically and rhythmically coherent music, offering a novel and efficient tool for both amateur and professional musicians. Additionally, image-to-music generation can raise ethical concerns, particularly around musical copyright issues. This work avoids such issues by not using deep learning in the music core. Although our methods are promising, current limitations include reduced lyrical coherence from image input and relatively simplistic music generation. Future work aims to enhance pitch construction by integrating chord progressions and cadences as well as exploring diverse musical genres for greater stylistic variety. These improvements will advance the ability of the framework to generate more contextually rich, musically complex, and genre-adaptive compositions, further positioning it as a creative copilot for music creation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *ArXiv Preprint ArXiv:2303.08774*, 2023. 1

[2] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 1

[3] Hangbo Bao et al. Neural melody composition from lyrics. *NLPCC*, 11838:499–511, 2019. 1

[4] K. Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024. 1

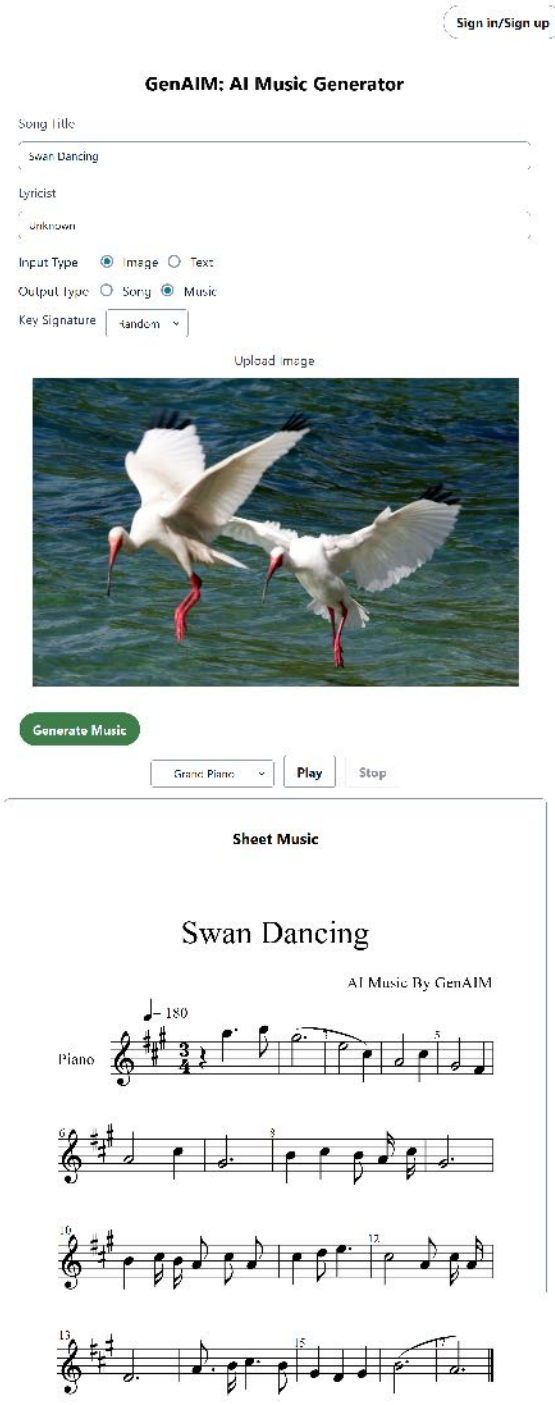

Figure 4. Image-to-Music Generation.

[5] Sanjoy Chowdhury, Sayan Nag, K J Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26816–26825, 2024. 2

[6] Sanjoy Chowdhury, Sayan Nag, K J Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26816–26825, 2024. 2

[7] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, pages 47704–47720. Curran Associates, Inc., 2023. 1

[8] Michael Scott Cuthbert and Christopher Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In *ISMIR*, pages 637–642. International Society for Music Information Retrieval, 2010. 2

[9] Hao-Wen Dong and Yi-Hsuan Yang. Convolutional generative adversarial networks with binary neurons for polyphonic music generation. *arXiv preprint arXiv:1804.09399*, 2018. 1

[10] Tanisha Hisariya, Huan Zhang, and Jinhua Liang. Bridging paintings and music – exploring emotion based music generation through paintings, 2024. 2

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. 1

[12] Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, and Yang Liu. Diff-BGM: A Diffusion Model for Video Background Music Generation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27338–27347, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

[13] Callie C. Liao, Duoduo Liao, and Jesse Guessford. Multimodal lyrics-rhythm matching. In *Proc. of the 2022 IEEE Int. Conf. on Big Data (BigData)*, pages 3622–3630, 2022. 1, 2, 3

[14] Callie C. Liao, Duoduo Liao, and Jesse Guessford. Automatic time signature determination for new scores using lyrics for latent rhythmic structure. In *Proc. of the 2023 IEEE Int. Conf. on Big Data (BigData)*, pages 4485–4494, 2023. 2

[15] Callie C. Liao, Duoduo Liao, and Ellie L. Zhang. Relationships between Keywords and Strong Beats in Lyrical Music . In *2024 IEEE International Conference on Big Data (BigData)*, pages 3191–3199, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1, 2, 3

[16] MIT. MIT Music21, 2025. 2

[17] Rohan Mitra and Imran Zualkernan. Music generation using deep learning and generative ai: A systematic review. *IEEE Access*, 13:18079–18106, 2025. 2

[18] Gwenaelle C. Sergio, Rammohan Mallipeddi, Jun-Su Kang, and Minho Lee. Generating music from an image. In *Proceedings of the 3rd International Conference on Human-Agent Interaction*, page 213–216, New York, NY, USA, 2015. Association for Computing Machinery. 2

[19] Yajie Wang, Mulin Chen, and Xuelong Li. Continuous emotion-based image-to-music generation. *IEEE Transactions on Multimedia*, 26:5670–5679, 2024. 2