Augmenting Human Creativity with Machine Learning

Hao-Wen Dong

University of Michigan
hwdong@umich.edu

Generative AI has been transforming the way we interact with technology and consume content. The recent successes of LLM-based chatbots and AI assistants have showcased how AI-powered technology can augment human creativity and boost human productivity. Generative AI technology will soon reshape how we create music, audio and video content in entertainment, commercial and educational fields, including the music, film, TV, podcast and gaming sectors.

Despite the successes of generative AI in certain fields, it has however been challenging to integrate generative AI into professional creative workflows for content creation for several reasons: 1) new application domains may require new generative models; 2) professionals need assistive tools that augment their creativity and productivity in addition to fully automated tools; 3) certain media require handling multimodal data streams at the same time.

My research aims to address these challenges and *augment human creativity with machine learning*. I develop human-centered generative AI technology that can be integrated into professional creative workflows, with a focus on music, audio, and video creation. My long-term goal is to lower the barrier of entry for content creation and democratize professional content creation for everyone.

In this talk, I will survey my representative work in the three main directions of my research: 1) generative models for music creation, 2) AI-assisted music creation tools, and 3) multimodal generative models for content creation.

Generative Models for Music Creation

First, I will talk about how I develop novel generative models for music creation. I will highlight my representative work on using generative adversarial networks (GANs) to generate multi-instrument music, which represents the first deep learning model for multitrack polyphonic music generation (Dong et al. 2018). I will also introduce my other contributions to the field in multitrack music generation, including follow-up work using GANs (Dong and Yang 2018), LSTMs (Dong et al. 2020), and transformers (Dong et al. 2023a; Ryu et al. 2024; Xu et al. 2025c). I also introduce our recent work on generating symbolic music from natural language prompts using an LLM-enhanced dataset (Xu et al. 2025c) and our recent work on extending a text-to-music model with a learnable video adapter (Kim et al. 2025).

AI-assisted Music Creation Tools

Second, I will talk about how I build AI-assisted content creation tools that aim to augment human creativity in their creative workflow. I will highlight my work on automatic music instrumentation that can produce convincing instrumentation for a solo piece (Dong et al. 2021). I will also briefly introduce my work on score-to-audio music performance synthesis (Dong et al. 2022) and our recent work on synthesizing expressive violin performance from sheet music (Kim, Dong, and Jeong 2025).

Multimodal Generative Models for Content Creation

Third, I will introduce my work on multimodal generative models for content creation that can process, understand, and generate data in multiple modalities at the same time. I will highlight our recent work on AI-assisted video editing, including a novel narration-centered documentary teaser generation system based on pretrained LLMs and language-vision models (Xu et al. 2025a). I will also talk about our recent work on long-to-short video editing where we proposed a novel retrieval-embedded generation framework that allows an LLM to quote multimodal resources while maintaining a coherent narrative (Xu et al. 2025b). I will also briefly introduce our work on self-supervised multimodal models that learn to separate and synthesize sounds from watching noisy videos for text-queried sound separation (Dong et al. 2023c) and text-to-audio synthesis (Dong et al. 2023b).

Future Work

Finally, I will conclude by discussing my future research directions on 1) multimodal generative AI for content creation, 2) human-AI co-creative tools for music, audio and video creation, and 3) human-like machine learning algorithms for music, movies and arts. In particular, I will highlight two ongoing research projects: First, I will discuss our future work towards next-generation video editing interfaces that can be integrated into the existing creative workflows of video creators using multimodal LLMs and retrieval embedded generation. Second, I will discuss our future work towards playful human-AI music co-creation systems where the user can *conduct* an AI orchestra that generates music on the fly through hand gestures and body movements.

References

Dong, H.-W.; Chen, K.; Dubnov, S.; McAuley, J.; and Berg-Kirkpatrick, T. 2023a. Multitrack Music Transformer. *ICASSP*.

Dong, H.-W.; Chen, K.; McAuley, J.; and Berg-Kirkpatrick, T. 2020. MusPy: A Toolkit for Symbolic Music Generation. *ISMIR*.

Dong, H.-W.; Donahue, C.; Berg-Kirkpatrick, T.; and McAuley, J. 2021. Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music. *ISMIR*.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018. MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. *AAAI*.

Dong, H.-W.; Liu, X.; Pons, J.; Bhattacharya, G.; Pascual, S.; Serrà, J.; Berg-Kirkpatrick, T.; and McAuley, J. 2023b. CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models. *WASPAA*.

Dong, H.-W.; Takahashi, N.; Mitsufuji, Y.; McAuley, J.; and Berg-Kirkpatrick, T. 2023c. CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos. *ICLR*.

Dong, H.-W.; and Yang, Y.-H. 2018. Training Generative Adversarial Networks with Binary Neurons by End-to-end Backpropagation. *arXiv preprint arXiv:1810.04714*.

Dong, H.-W.; Zhou, C.; Berg-Kirkpatrick, T.; and McAuley, J. 2022. Deep Performer: Score-to-Audio Music Performance Synthesis. *ICASSP*.

Kim, D.; Dong, H.-W.; and Jeong, D. 2025. ViolinDiff: Enhancing Expressive Violin Synthesis with Pitch Bend Conditioning. *ICASSP*.

Kim, H.; Novack, Z.; Xu, W.; McAuley, J.; and Dong, H.-W. 2025. Video-Guided Text-to-Music Generation Using Public Domain Movie Collections. *ISMIR*.

Ryu, J.; Dong, H.-W.; Jung, J.; and Jeong, D. 2024. Nested Music Transformer: Sequentially Decoding Compound Tokens in Symbolic Music and Audio Generation. *ISMIR*.

Xu, W.; Liang, P. P.; Kim, H.; McAuley, J.; Berg-Kirkpatrick, T.; and Dong, H.-W. 2025a. TeaserGen: Towards Generating Teasers for Long Documentaries. *ICLR*.

Xu, W.; Ma, Y.; Huang, J.; Li, Y.; Ma, W.; Berg-Kirkpatrick, T.; McAuley, J.; Liang, P. P.; and Dong, H.-W. 2025b. REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing. *NeurIPS*.

Xu, W.; McAuley, J.; Berg-Kirkpatrick, T.; Dubnov, S.; and Dong, H.-W. 2025c. Generating Symbolic Music from Natural Language Prompts using an LLM-Enhanced Dataset. *ISMIR*.