MIDILM: A Dual-Path Model for Controllable Text-to-MIDI Generation

Shuyu Li[1*], **Dooho Choi**[1*], **Yunsick Sung**[1†]

[1]Department of Computer Science & Artificial Intelligence, Dongguk University-Seoul, Seoul, 04620, Korea
lishuyu@dongguk.edu, likeb789@dgu.ac.kr, sung@dongguk.edu

Abstract

Text-to-MIDI generation offers editable and hierarchical control over symbolic music generation. Previous approaches either convert text into a limited set of musical attributes and generate music based on these attributes, which limits semantic controllability, or use end-to-end models that map text directly to music without deeply aligning the features of both modalities, often resulting in a lack of structural coherence and mismatches in key, meter, and tempo. We propose MIDILM, which addresses these limitations by employing text conditioning with a dual-path decoder that processes textual and musical information through separate feedforward paths following a shared masked self-attention mechanism. On the MidiCaps benchmark, MIDILM outperformed the strongest baseline, with relative improvements ranging from 6.07% on CLAP to 144.77% on TB across semantic alignment and structural metrics. These gains confirm its ability to enhance both semantic controllability and structural coherence. Collectively, we expect that MIDILM will serve as a useful reference framework for future investigations into controllable and structurally faithful cross-modal music generation.

Code — https://github.com/Large-Multimodal-Model-Lab/MIDILM

Introduction

The rapid advancement of AI-generated content has spurred extensive research into cross-modal generation (Baltrušaitis, Ahuja, and Morency 2018; Tsimpoukelli et al. 2021). Significant progress has already been made in areas such as text-to-image (Rombach et al. 2022; Saharia et al. 2022) and text-to-speech (Shen et al. 2018; Radford et al. 2023) generation. In text-to-music research, audio music and symbolic music are regarded as two distinct domains. In this paper, we focus on the symbolic music domain, aiming to generate Musical Instrument Digital Interface (MIDI) sequences that accurately reflect the semantics of natural language inputs.

MIDI encodes musical elements such as pitch, duration, velocity, and instrumentation as discrete events, offering precise control, transparent structure, and flexible post-editing capabilities. While generating high-quality audio music directly from text has become increasingly feasible (Agostinelli et al. 2023; Copet et al. 2023; Liu et al. 2024; Chen et al. 2024), the research that converts free-form text into symbolic music in the MIDI format accurately still faces several challenges (Lu et al. 2023; Bhandari et al. 2025; Roy, Puri, and Herremans 2025). First, there exists a fundamental difference between modalities, as text communicates high-level semantics, emotional nuance, and structural details, whereas MIDI encodes music as an event-based timeline that requires explicit coordination across musical dimensions. Second, MIDI exhibits intrinsic structural

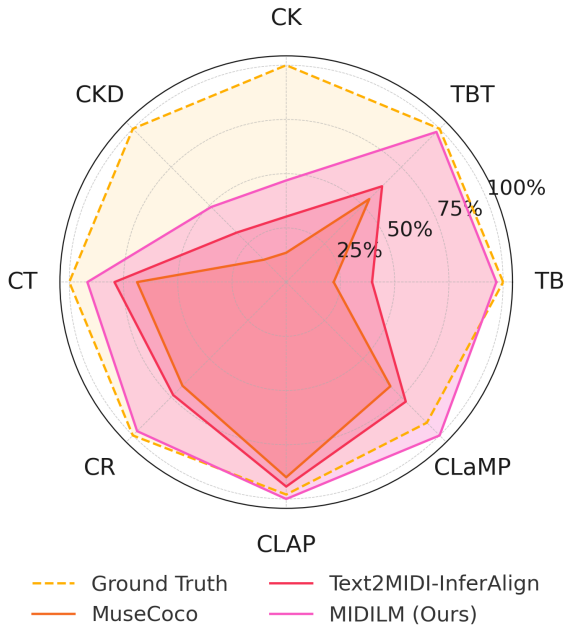

Figure 1: Benchmark comparison on the MidiCaps test set for MIDILM, Text2MIDI-InferAlign, and ground truth. Metrics include TB, TBT, CR, CKD, and CK (Bhandari et al. 2025; Roy, Puri, and Herremans 2025), as well as CLAP (Wu et al. 2023) and CLaMP 3 (Wu et al. 2025). All values are normalized to the range 0 to 100%. A larger area indicates better performance.

[*]These authors contributed equally.
[†]Corresponding author. Email: sung@dongguk.edu.

complexity and strict organization, as it consists of diverse and interdependent event types such as pitch, velocity, duration, meter, and instrumentation, all of which must be precisely arranged to produce coherent musical output. Successful generation must not only ensure semantic alignment but also maintain structural coherence.

To narrow this gap, researchers have proposed attribute pipelines and direct sequence models. MuseCoco (Lu et al. 2023) predicts a fixed set of musical attributes from text and then generates MIDI conditioned on them. However, the rigid attribute space limits nuance and prohibits fine adjustments once the attributes are chosen. Text2MIDI (Bhandari et al. 2025) leverages a FLAN-T5 (Chung et al. 2024) encoder and a Transformer decoder, integrated via cross-attention. Although this design enhances overall text-MIDI alignment, the degree of alignment remains insufficient. Text2MIDI-InferAlign (Roy, Puri, and Herremans 2025), as an advanced version of Text2MIDI, introduces reward shaping at inference time to enhance text-MIDI alignment. However, because it applies adjustments only after generation, it remains limited in accurately capturing structural details such as key and tempo. These works have laid an important foundation for the field, but substantial challenges remain, especially in achieving both semantic controllability and structural coherence.

We propose MIDILM, a text-driven model that simultaneously enables precise semantic control and structurally faithful MIDI generation. For text-conditioned generation, MIDILM employs a dual-path architecture in which text tokens and MIDI tokens are input together but processed through separate computational paths. The text tokens are handled by a dedicated multilayer perceptron (MLP) for language semantics, while the MIDI tokens are processed by Mixture-of-Experts (MoE) (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022; Rajbhandari et al. 2021; Touvron et al. 2023; Jiang et al. 2024) responsible for modeling musical structure. This design significantly strengthens both semantic controllability and structural coherence in generated music.

The main contributions of this paper are as follows:

- Unlike modality-fused techniques, we model textual semantics and musical structure in dedicated paths, integrating them only through a shared masked self-attention layer. We refer to this design principle as modality isolation, which keeps modality-specific transformations independent while allowing interaction solely via the shared attention mechanism.
- MIDILM introduces a dual-path decoder architecture that combines a MoE module for musical structure modeling with a dedicated MLP path for language semantics, enabling efficient and scalable generation.
- Experiments on structured music generation tasks demonstrate that MIDILM achieves state-of-the-art results, surpassing strong baselines in semantic alignment and structural accuracy. The results, as shown in Figure 1, confirm the effectiveness of the proposed design.

Related Work

Recent advances in symbolic music generation have capitalized on natural language to produce MIDI content, supporting personalized composition, automatic arrangement, and interactive tools. Challenges persist in achieving precise semantic alignment and ensuring structural coherence.

MuseCoco (Lu et al. 2023) frames text-to-MIDI generation as a two-stage process. It first predicts musical attributes from text and then converts these attributes into music. The explicit attribute layer provides controllability. However, its fixed vocabulary limits expressiveness, as it cannot fully capture the nuances embedded in unrestricted language. As a result, stylistic diversity and creative fidelity are constrained.

Text2MIDI (Bhandari et al. 2025) proposes an end-to-end model capable of directly generating high-quality MIDI files from textual descriptions. The approach employs a pre-trained FLAN-T5 as the encoder, which embeds the input text and conditionally feeds it into a Transformer decoder to autoregressively generate symbolic music sequences in the REMI+ (Huang and Yang 2020; von Rütte et al. 2022; Mittal et al. 2021) representation. The model is first semi-supervised pre-trained on the SymphonyNet (Liu et al. 2022) dataset, where pseudo-text descriptions are automatically extracted from MIDI attributes, and then fine-tuned on the large-scale MidiCaps (Melechovsky, Roy, and Herremans 2024) dataset with paired text-MIDI data. Compared to MuseCoco, the end-to-end modeling approach of Text2MIDI substantially improves expressive capability.

Text2MIDI-InferAlign (Roy, Puri, and Herremans 2025) improves Text2MIDI through reward shaping applied at inference time. The approach introduces objectives for text-audio consistency, measured by the CLAP (Wu et al. 2023) score, and for harmonic conformity, penalizing notes that conflict with the intended key. Although this strategy enhances alignment without additional training, it cannot fully address expressiveness limitations. Complex tempo, key, and meter cues remain inconsistently realized because the underlying representations are untouched.

Inspired by the end-to-end paradigm of Text2MIDI-InferAlign, we introduce MIDILM with two improvements. First, we replace cross-attention with prefix conditioning to stabilize optimization and reduce both parameters and computation. Second, we exploit the categories of MIDI tokens and use a MoE router that activates only a few experts per step to allocate capacity and limit interference. This yields stronger alignment, better structural coherence, and richer expressiveness at lower cost. We further adopt a dual-path design with modality isolation, where the text path processes text tokens and the MIDI path processes MIDI tokens, with both paths coupled through shared self-attention.

Method

Architecture

Figure 2 provides an overview of the model architecture, where each decoder layer applies RMSNorm (Zhang and Sennrich 2019) before both the shared masked self-attention stack and the modality-specific feedforward paths.

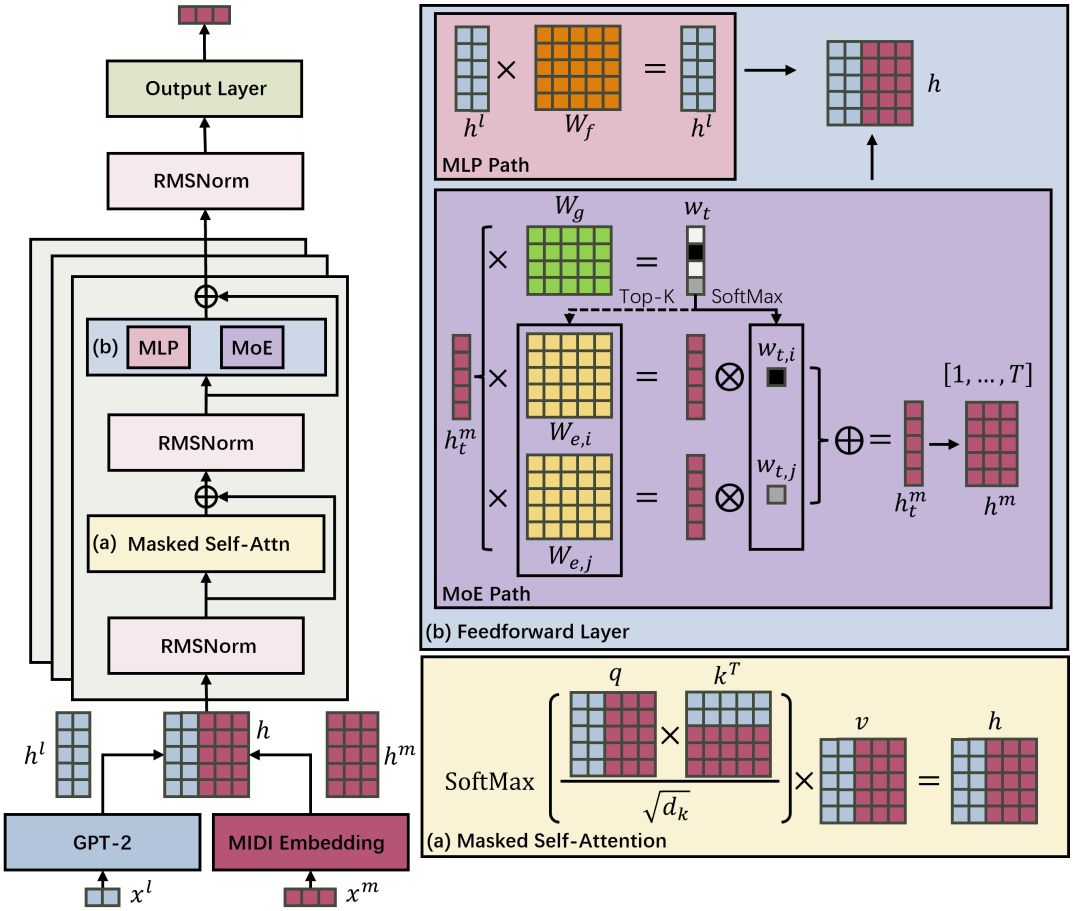

Figure 2: Architecture of the proposed MIDILM. (a) The masked self-attention mechanism processes the concatenated text tokens and MIDI tokens. (b) The feedforward layer contains separate MLP and MoE paths that are responsible for processing text tokens and MIDI tokens, respectively.

The text tokens embedded using a pretrained GPT-2 (Radford et al. 2019) model are concatenated with the embedded MIDI tokens along the time axis and passed into the same stack of decoder blocks. Masked self-attention (Figure 2(a)) enables every MIDI position to attend to the entire text prefix, ensuring that global semantic information is accessible at all decoding steps, while preserving the autoregressive order of musical events.

In the feedforward layer (Figure 2(b)), the architecture adopts a dual-path design:

- **MLP path** processes text hidden states h^l; the module W_f transforms them into an updated h^l.
- **MoE path** processes MIDI hidden states in token-level h_t^m; the router W_g computes gating logits and applies SoftMax with Top-K selection to obtain sparse weights $w_{t,i}$ and $w_{t,j}$, which activate the selected experts $W_{e,i}$ and $W_{e,j}$. The weighted sum at the token-level for each time step yields h_t^m, and concatenation of all h_t^m for $t = 1, \ldots, T$ produces the updated h^m.

Where W_f and each $W_{e,*}$ are implemented as three-layer structures comprising an up-projection, an intermediate layer, and a down-projection, and are shown in simplified form in the figure.

After separate processing, the updated h^l and h^m are concatenated to form the hidden state h. After all layers, only h^m from the final layer is fed into output normalization and projection to produce the MIDI token logits.

Training

The optimization of MIDILM combines a token-level likelihood term with a factor that regularizes expert routing.

Cross-entropy Term The model is optimized by minimizing the cross-entropy loss, where x_t^m denotes the MIDI token at time step t, $x_{<t}^m$ refers to all previously generated MIDI tokens, and x^l is the input text token sequence.

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P\left(x_t^m \mid x_{<t}^m, x^l\right) \tag{1}$$

MoE Load-balancing Term Without regulation, a small subset of experts tends to dominate routing. Following (Shazeer et al. 2017), we measure the fraction of tokens sent

to each expert in a mini-batch (vector $\mathbf{n} \in \mathbb{R}^E$) and penalize the squared coefficient of variation:

$$\mathcal{L}_{\text{LB}} = \left(\frac{\text{std}(\mathbf{n})}{\text{mean}(\mathbf{n})}\right)^2 \quad (2)$$

Combined Loss The total training objective is defined as a weighted sum of the cross-entropy and the load-balancing term:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{LB}} \quad (3)$$

Here, the scalar λ is a fixed hyperparameter that controls the strength of the load-balancing regularization.

Inference

During inference, the model first encodes the natural language text into the decoder space and initializes generation with the special `[BOS]` token. The decoder then autoregressively produces MIDI tokens, reusing the key–value cache at each step and sampling via a temperature-controlled Gumbel-Softmax, until the `[EOS]` token appears. The resulting REMI+ token sequence is then converted into a standard MIDI file.

Experiment

Dataset and Representation

We conducted all experiments on the MidiCaps (Melechovsky, Roy, and Herremans 2024) dataset, which consists of 168,385 MIDI music files paired with natural language descriptions. All original MIDI files were sourced from the Lakh MIDI Dataset (Raffel 2016). We followed the official 9:1 train/test split of MidiCaps. The official training set was kept unchanged, and the official test set was randomly divided into two equal halves, using 50% as a validation set and the remaining 50% as our held-out test set.

In addition, we constructed a free-form text evaluation set to assess model generalization under open-ended scenarios. We manually curated 96 texts to cover diverse genres, emotions, and instrumentation, with each text ranging from 10 to 20 words. For example, a representative text is: *"A heartfelt pop ballad expressing longing, featuring piano, electric guitar, and soft strings."*

We adopted distinct representation strategies for the MIDI and text modalities. For MIDI, we used the REMI+ representation (Huang and Yang 2020; von Rütte et al. 2022; Mittal et al. 2021) and performed tokenization with `miditok` (Fradet et al. 2023), resulting in a vocabulary of 515 tokens. Text inputs were tokenized using the GPT-2 tokenizer (Radford et al. 2019), which comprises 50,257 tokens.

Experimental Setup

Model Configuration MIDILM employed a 12-layer unidirectional Transformer decoder with rotary positional encoding (Su et al. 2024) and a local MoE design, where each layer contained 8 experts and applied top-2 routing for each token. A dedicated, higher-capacity MLP path was used to process text tokens, while the main MoE path handled MIDI tokens.

Training Details MIDILM was trained on the MidiCaps training set for a total of 100,000 steps, using a cosine decay learning rate schedule with 10,000 warmup steps. The learning rate peaked at 1×10^{-4} and decayed to 1×10^{-6}. With a batch size of 16, each batch consisted of MIDI sequences of length 2,048 and corresponding text sequences of length 128. Training was conducted on four NVIDIA RTX 3090 GPUs and completed in approximately 60 hours.

Baselines We compared MIDILM with two representative baselines including Text2MIDI-InferAlign (Roy, Puri, and Herremans 2025) and MuseCoco (Lu et al. 2023). All models used a sampling temperature of 1.0 when generating evaluation samples.

Objective Metrics We evaluated model performance using the standard metrics employed in Text2MIDI-InferAlign (Roy, Puri, and Herremans 2025). These metrics include:

- **Compression Ratio (CR)**: A structural compression ratio computed using the COSIATEC algorithm (Meredith 2013) to quantify repeated patterns and long-term structures in the MIDI sequence.
- **CLAP** (Wu et al. 2023): The cross-modal similarity between the generated audio (converted from MIDI) and the input text, as measured by the CLAP model. For audio conversion, we used the unified instrument library. The CLAP evaluation was conducted with the `music_audioset_epoch_15_esc_90.14.pt` checkpoint.
- **Tempo Bin (TB)**: The proportion of samples whose extracted tempo fell within the correct predefined tempo bin.
- **Tempo Bin with Tolerance (TBT)**: The proportion of samples where the predicted tempo was within the ground truth bin or an adjacent bin.
- **Correct Key (CK)**: The proportion of samples where the predicted key matched the ground truth.
- **Correct Key with Duplicates (CKD)**: The proportion of samples where the predicted key matched the ground truth or a duplicate key.

To further evaluate performance in terms of time-signature consistency and text-MIDI alignment, we introduced two additional metrics:

- **Correct Time-signature (CT)**: The proportion of generated MIDI files in which the predicted meter exactly matched the ground truth, thus capturing structural accuracy not addressed by prior benchmarks.
- **CLaMP 3** (Wu et al. 2025): Directly measures cross-modal semantic alignment between the generated MIDI and the input text without converting MIDI into audio, providing a detailed assessment of semantic consistency between natural language and symbolic music.

Listening Test Subjective musical quality was assessed through a blind listening test with 15 raters, consisting of 5 trained musicians and 10 non-musicians. All raters evaluated each sample in a blind setting, without information

Metric	Ground Truth	MuseCoco	Text2MIDI-InferAlign	MIDILM (Ours)	w/o All	w/o ModIso	w/o MoE
TB ↑	N/A	0.2188	0.3958	0.9688	0.8269	0.8694	0.9132
TBT ↑	N/A	0.5417	0.6250	0.9792	0.8675	0.9046	0.9363
CK ↑	N/A	0.1354	0.3021	0.4688	0.3834	0.4159	0.4426
CKD ↑	N/A	0.1458	0.3229	0.4896	0.3967	0.4278	0.4549
CT ↑	N/A	0.6875	0.7917	0.9167	0.8358	0.8773	0.8968
CR ↑	2.5327	1.7123	1.8658	2.4617	2.2139	2.3244	2.3993
CLAP ↑	0.3063	0.2814	0.2948	0.3127	0.2867	0.2993	0.3054
CLaMP 3 ↑	0.1434	0.1062	0.1219	0.1564	0.1316	0.1405	0.1478

Table 1: Objective evaluation on the MidiCaps test set. Besides baseline comparisons, three ablation variants (w/o All, w/o ModIso, w/o MoE) are included to quantify the impact of modality isolation and MIDI-side MoE routing. MIDILM consistently outperforms all baselines and ablations (Welch t-tests, $p < 0.0001$). All results are averaged over 10 different random seeds, each corresponding to a randomly sampled subset of 96 test instances drawn from the test set pool.

about which model produced it. The Mean Opinion Score (MOS) was used as the evaluation metric. For each sample, the 15 raters independently assigned a score from 1 to 5 considering musicality, coherence, and relevance to the text. The MOS scale was defined as follows:

1: Unrecognizable as music or entirely irrelevant to the text.

2: Barely musical, low coherence, or weak text alignment.

3: Some musicality and partial text relevance, but contained clear flaws.

4: Mostly musical, coherent, and relevant to the text, with minor imperfections.

5: Highly musical, coherent, and fully aligned with the text.

The final MOS for each system was computed as the mean over the 15 raters and all evaluated samples. For each system, including ground truth, Text2MIDI-InferAlign, and MIDILM, 96 samples from the test set and 96 samples from the free-form text set were evaluated.

Results and Analysis

Objective Evaluation Table 1 shows a clear structural and semantic advantage for MIDILM. TB increased substantially, and TBT followed the same trend, which indicated that the model selected tempos that matched the text rather than repeating the corpus dominant range. CT also improved, and this was attributed to the ability of the dual path decoder to represent rare meters, whereas the baselines mostly produced the corpus dominant *4/4*. Although CK and CKD improved over the baselines, substantial room for progress remained. The primary limitation was that standard MIDI does not provide explicit key-related tokens, forcing the model to infer key information solely from global pitch distributions. Cross-modal alignment scores measured by CLAP and CLaMP 3 increased substantially and even surpassed the ground truth, suggesting that the generated music captured textual semantics more faithfully. This is likely because the model is explicitly optimized for semantic consistency, while certain ground-truth samples contain weaker text-MIDI alignment due to subjective annotations or dataset noise, allowing the generated outputs to achieve higher scores. For CR, MIDILM scored between the baselines and the ground truth, which suggested that its gains were not achieved through redundant padding. Ablation results confirm the contribution of each component. Removing both yields the largest degradation, highlighting their complementarity. Without modality isolation, text and MIDI tokens interfere within shared experts, and without MoE Routing, MIDI specialization weakens, reducing musical expressiveness.

Table 2 further shows that free-form texts, which lacked explicit constraints on tempo, meter, and key, reduced absolute alignment scores and narrowed the gaps among systems. Even so, MIDILM maintained a modest edge on CLAP and CLaMP 3 and achieved the highest CR, which indicated that it generated longer and information-dense passages without padding and adapted flexibly to different input texts.

Metric	MuseCoco	Text2MIDI-InferAlign	MIDILM (Ours)
CR ↑	1.9652	2.1414	3.1102
CLAP ↑	0.1767	0.1847	0.1881
CLaMP 3 ↑	0.1366	0.1326	0.1390

Table 2: Objective evaluation results on free-form texts.

Subjective Evaluation Table 3 shows that listeners were able to perceive the structural improvements introduced by the model. On the test set, the MOS for MIDILM exceeded Text2MIDI-InferAlign by nearly one point. In the free-form text setting, the score margin became smaller, as evaluators placed greater emphasis on fluency and emotional expressiveness rather than strict adherence to structural features. Nonetheless, MIDILM continued to obtain comparatively high MOS values, attributable to its smoother phrasing and richer instrumentation.

Qualitative Results and Visualizations

MoE Routing and Prefix Attention Figure 3 illustrates expert specialization in the MoE architecture. For each token, layer-wise routing probabilities were concatenated to form a high-dimensional expert-assignment fingerprint, and all fingerprints were stacked into a global expert-utilization matrix. Principal component analysis (PCA) projected this matrix to two dimensions, so that Euclidean distance reflected routing similarity. Tokens were color-coded by mu-

Metric	Ground Truth	MuseCoco	Text2MIDI-InferAlign	MIDILM (Ours)
Test (MOS±std) ↑	4.1010±0.9590	2.0854±0.9499	2.4861±0.7675	3.2306±0.9160
Free (MOS±std) ↑	N/A	3.0312±0.4886	3.2965±0.6025	3.3417±0.5976

Table 3: MOS evaluation results on MidiCaps test set and free-form texts.

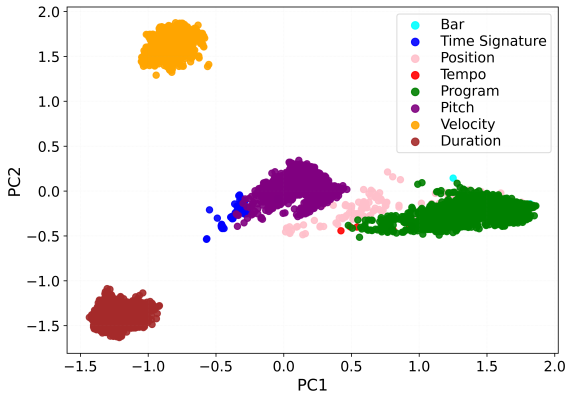

Figure 3: PCA projection of MoE expert routing fingerprints for each token type. Each point represents the expert activation pattern for a token across layers. The X and Y axes correspond to the first two principal components (PC1 and PC2). Clusters reflect routing similarity between token types based on expert assignment patterns.

sical category; tight same-color clusters indicated that the MoE assigned functionally related tokens to the same or similar experts. In the figure, *Velocity* was clearly separated from other categories, showing that dynamic intensity was managed by dedicated experts. *Duration* also formed an isolated region, suggesting effective disentanglement of the rhythmic backbone to avoid interference. In contrast, *Time Signature*, *Pitch*, *Position*, *Tempo*, *Program*, and *Bar* were grouped closely, reflecting their frequent collaboration in melodic organization, instrumentation, and tempo marking. Overall, the MoE automatically disentangled the most critical dimensions for performance dynamics and rhythm (*Velocity* and *Duration*), while maintaining strong correlations among the remaining features to support coherent musical structure.

Figure 4 presents the self-attention weights of the 12-layer decoder. The horizontal axis represented the source sequence, which included both text tokens and MIDI tokens, while the vertical axis labeled Predict indicated the current prediction position from top to bottom. Each subplot corresponded to a decoder layer, with a red dividing line marking the boundary between text and MIDI tokens. In Layers 1 through 8, the attention mainly focused along the main diagonal, reflecting the ability of the model to capture local autoregressive dependencies. As the layers deepened, particularly from Layer 9 to Layer 12, distinct vertical stripes emerged to the left of the dividing line, demonstrating a strong focus on the text prefix. Notably, as the network went deeper, the influence of global text semantics not only per-

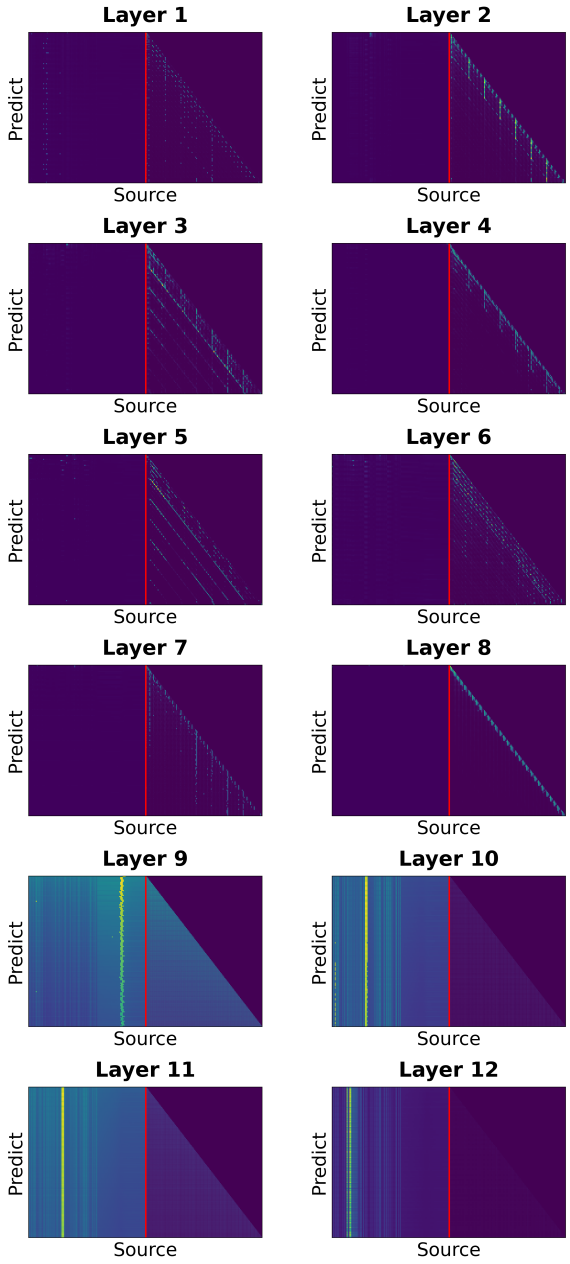

Figure 4: Visualization of self-attention weights across 12 decoder layers. The X axis shows input tokens, and the Y axis shows predicted tokens. Text tokens appear to the left of the red line, MIDI tokens to the right. Bright colors indicate strong attention, while dark colors indicate weak attention.

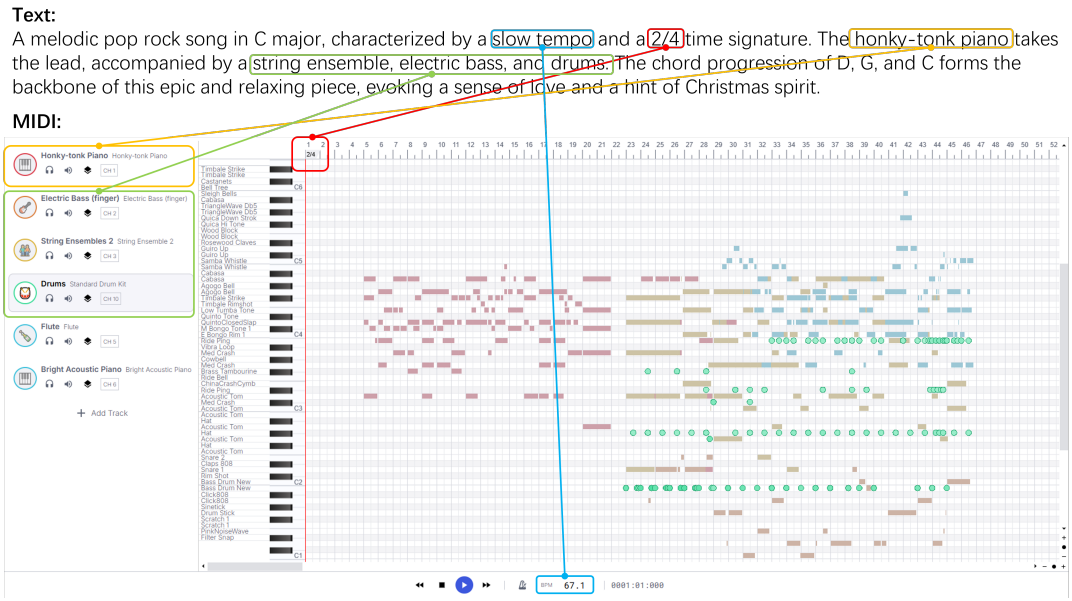

Figure 5: Text-to-MIDI generation example by MIDILM. The input text describes a slow-tempo pop rock piece in C major with a 2/4 meter, led by honky-tonk piano and supported by strings, electric bass, and drums. The generated MIDI matches these specifications: the piano leads (yellow), the meter is 2/4 (red), and the tempo is 67 BPM (blue). The presence of strings, electric bass, and drums (green) further confirms accurate instrumentation alignment, with colored overlays indicating how textual cues map to structural elements in the output.

sisted but became even more pronounced, continuously providing essential guidance for the music generation process.

Text-to-MIDI Case Study Figure 5 presents a MIDI generation example based on an input text, comparing the outputs from MIDILM and the ground truth sample. The figure included the input text and visualizations of the generated MIDI from MIDILM, rendered with `signal`. The results showed that MIDILM achieved high consistency with the semantic requirements of the text in terms of instrumentation, meter, and tempo.

Discussion

MIDILM supports two training paradigms and aligns heterogeneous modalities through a unified dual-path architecture. In the from-scratch setting, the model is trained on paired text-MIDI data, allowing shared self-attention and dual-path feed-forward layers to co-evolve and learn both cross-modal coupling and modality-specific representations without relying on pretrained checkpoints. As a fine-tuning method, the architecture adapts flexibly. For music-side adaptation, the text path and shared attention remain frozen while the MoE path and output projection are updated. For text-side adaptation, the MIDI path is frozen while the text path and projection layer are updated. When stronger cross-modal alignment is needed, lightweight adapters or low-rank updates can be inserted into the shared attention so that most parameters stay fixed. This design enables efficient task adaptation while minimizing the risk of catastrophic forgetting.

Conclusion

MIDILM introduces a shared-attention, dual-path feed-forward architecture that effectively decouples yet deeply fuses textual semantics with symbolic music structure. On the MidiCaps benchmark, it surpassed the strongest baseline, Text2MIDI-InferAlign, with improvements ranging from 6.07% to 144.77% across structural and alignment metrics. MOS tests further confirmed a pronounced improvement in perceived musical quality. Routing and attention visualizations revealed that Mixture-of-Experts modules self-organized, with specialist experts focusing on velocity, duration, and other salient dimensions, while deeper layers persistently reinforced the textual prefix, explaining the gains in rhythmic control and emotional expression. Nevertheless, MIDILM still underperforms in key alignment, and its output quality declines for pieces with complex tonality or frequent modulations. Overall, the framework advances semantic controllability and structural coherence while providing an interpretable foundation for cross-modal music generation, paving the way for future research that explicitly encodes key and rhythmic cues, employs retrieval-augmented hierarchical generation, and integrates real-time human-AI interaction with multidimensional subjective evaluation protocols.

Acknowledgments

This research was supported by the Strengthening AI Computing Resource Utilization Foundation, funded by the Government of the Republic of Korea (Ministry of Science and ICT) (RQT-25-090179). This work was sup-

ported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16068981). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00254592) grant funded by the Korea government (MSIT).

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*.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Bhandari, K.; Roy, A.; Wang, K.; Puri, G.; Colton, S.; and Herremans, D. 2025. Text2midi: Generating symbolic music from captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23478–23486.

Chen, K.; Wu, Y.; Liu, H.; Nezhurina, M.; Berg-Kirkpatrick, T.; and Dubnov, S. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1206–1210. IEEE.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36: 47704–47720.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fradet, N.; Briot, J.-P.; Chhel, F.; Seghrouchni, A. E. F.; and Gutowski, N. 2023. MidiTok: A python package for MIDI file tokenization. *arXiv preprint arXiv:2310.17202*.

Huang, Y.-S.; and Yang, Y.-H. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM international conference on multimedia*, 1180–1188.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2871–2883.

Liu, J.; Dong, Y.; Cheng, Z.; Zhang, X.; Li, X.; Yu, F.; and Sun, M. 2022. Symphony generation with permutation invariant language model. *arXiv preprint arXiv:2205.05448*.

Lu, P.; Xu, X.; Kang, C.; Yu, B.; Xing, C.; Tan, X.; and Bian, J. 2023. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*.

Melechovsky, J.; Roy, A.; and Herremans, D. 2024. MidiCaps: A large-scale MIDI dataset with text captions. *arXiv preprint arXiv:2406.02255*.

Meredith, D. 2013. COSIATEC and SIATECCompress: Pattern discovery by geometric compression. In *International society for music information retrieval conference*. International Society for Music Information Retrieval.

Mittal, G.; Engel, J.; Hawthorne, C.; and Simon, I. 2021. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C. 2016. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Columbia University.

Rajbhandari, S.; Ruwase, O.; Rasley, J.; Smith, S.; and He, Y. 2021. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, 1–14.

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.

Roy, A.; Puri, G.; and Herremans, D. 2025. Text2midi-InferAlign: Improving Symbolic Music Generation with Inference-Time Alignment. *arXiv preprint arXiv:2505.12669*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779–4783. IEEE.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

von Rütte, D.; Biggio, L.; Kilcher, Y.; and Hofmann, T. 2022. Figaro: Generating symbolic music with fine-grained artistic control. *arXiv preprint arXiv:2201.10936*.

Wu, S.; Zhancheng, G.; Yuan, R.; Jiang, J.; Doh, S.; Xia, G.; Nam, J.; Li, X.; Yu, F.; and Sun, M. 2025. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2605–2625.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhang, B.; and Sennrich, R. 2019. Root mean square layer normalization. *Advances in neural information processing systems*, 32.