ConvM2D2: Improving Generative Music Evaluation using Self-Supervised Alternative to CLAP

Kehinde Elelu[1,*], Josh Siegel[1,*], Ali Saffary[1], Hung Luong[1], Simeon Babatunde, Ebuka Okpala

***Abstract*—Evaluating the perceptual quality of AI-generative music remains a challenge in music information retrieval and computational creativity applications. Approaches such as those adopted in the MusicEval and AudioMOS challenges primarily rely on CLAP, a contrastive audio-text model to extract embeddings for Mean Opinion Score (MOS) prediction. While CLAP excels at coarse audio-text alignment, it struggles to capture fine-grained musical attributes such as timbral richness, rhythmic precision, and structural coherence, leading to suboptimal alignment with expert human evaluations. We introduce ConvM2D2, a novel dual-branch neural architecture that leverages M2D2, a second-generation masked modeling framework, as the upstream audio encoder for MOS prediction. M2D2 is trained to reconstruct masked audio segments, enabling it to capture temporally- and acoustically-detailed features that more closely reflect human perceptual criteria. The ConvM2D2 model processes audio and text embeddings jointly through specialized convolutional and multi-layer perceptron pathways to predict both Overall Musical Quality and Textual Alignment scores. We evaluate ConvM2D2 on the MusicEval benchmark, comparing its performance against other models and achieve improvements across all evaluation metrics (MSE, LCC, SRCC, and KTAU) at both utterance- and system-level evaluation. ConvM2D2 reaches a system-level LCC of 0.964 and reduces MSE by 88% compared to the baseline, demonstrating strong alignment with human judgments across both overall musical quality and textual alignment tasks. This big improvement indicates ConvM2D2 can judge AI-generated music much more like a musical expert, making it easier to find, improve, and recommend better-sounding music.**

***Index Terms*—AI-generated music, perceptual quality, music evaluation, CLAP, M2D2, ConvM2D2, contrastive learning, masked modeling, audio-text alignment, mean opinion score (MOS), MusicEval, AudioMOS, music information retrieval, computational creativity.**

I. Introduction

A surge in generative models for music synthesis has opened new avenues for creative uses of AI, enabling systems to produce increasingly convincing and diverse musical outputs [1], [2]. As illustrated in Figure 1, a textual prompt is provided as input to a Text-to-Music system, which then generates a large-scale dataset of audio samples conditioned on the prompt. However, evaluating the perceptual quality of these generative outputs remains a bottleneck. Unlike objective tasks such as transcription or source separation, music quality assessment involves subjective, multi-dimensional human judgments that are challenging to quantify algorithmically [3]. As of July 2025, human evaluations via Mean Opinion Scores (MOS) remain the gold standard [4], but these are labor-intensive, costly, and inherently non-scalable. This motivates the need for automated, perceptually aligned evaluation models that can serve as reliable proxies for expert human ratings. [5].

Manuscript received July 10, 2021

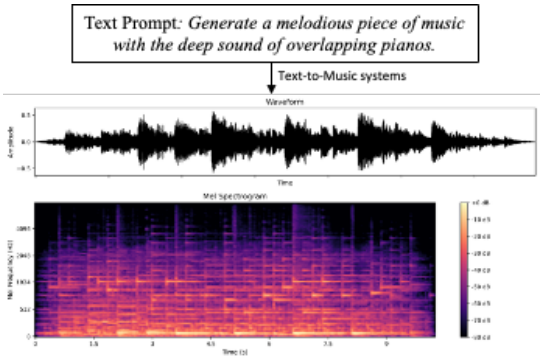

Fig. 1. Generative music AI models leverage textual prompts to synthesize audio music samples, advancing tasks in music information retrieval and computational creativity. The MusicEval dataset, containing diverse prompts and model-generated outputs, is used to evaluate the quality and alignment of the generated music.

Prior efforts, including the MusicEval and AudioMOS challenges, primarily employ Contrastive Language-Audio Pretraining (CLAP), a contrastive audio-text pretrained model, to extract embeddings for MOS prediction [6]. While CLAP effectively models cross-modal alignment, its training objective prioritizes broad audio-text similarity rather than capturing intricate musical properties such as timbral texture, rhythmic accuracy, melodic coherence, and dynamic variation [7]. As a result, existing CLAP-based MOS prediction systems often struggle to align with the evaluations provided by expert listeners.

In this work, we introduce ConvM2D2, a new architecture designed to address these limitations by leveraging the Masked Modeling Duo (M2D2) framework as the upstream audio encoder for generative music evaluation. Unlike CLAP, M2D2 employs self-supervised masked reconstruction objectives that force the model to learn fine-grained temporal and acoustic features that are critical to human perceptions of musical quality [8]. This appears to be the first time M2D2 is being leveraged for subjective music evaluation tasks.

This paper presents several contributions to the field of automatic evaluation of AI-generated music, particularly in the context of perceptual quality assessment:

- **We introduce a new dual-branch ConvM2D2 model** that uses both audio and text features to separately predict scores for Overall Musical Quality and Textual Alignment.
- **We thoroughly test ConvM2D2 on the MusicEval**

dataset and show that it beats both the official challenge baseline and the top-performing WhisQw/OT model [9] on all major metrics, including **MSE, LCC, SRCC, and KTAU**.

- **We show that using M2D2's masked modeling for audio** creates richer audio features, which match expert human opinions about music quality much better than the older CLAP model.

To our knowledge, this represents the first application of ConvM2D2 for generative music evaluation, and this approach supports a more reliable automated evaluation framework for advancing generative music systems. The development of ConvM2D2 for generative music evaluation has the potential to democratize access to high-quality music assessment tools. Our approach enables researchers and creators from diverse backgrounds to evaluate generative music systems more efficiently. This work paves the way for more inclusive, scalable, and transparent evaluation pipelines in computational creativity and music AI.

II. Background and Related Works

A. Generative Music Quality Evaluation

Evaluating the perceptual quality of AI-generated music remains a challenging area of research in music information retrieval and computational creativity [10], [11]. Traditional evaluation metrics such as Inception Score (IS) and Fréchet Audio Distance (FAD) have been adapted from generative image models to assess distributional similarity between generated and real music datasets [12], [13]. However, these metrics are often poorly aligned with subjective human judgment, particularly for complex musical dimensions such as expressiveness, coherence, and semantic alignment with conditional inputs.

To address these limitations, human evaluations using Mean Opinion Scores (MOS) remain the gold standard for assessing generative music quality. MOS assessments involve collecting ratings from expert or lay listeners on a Likert scale, providing direct insight into perceived quality and relevance. Recognizing the need for scalable and reproducible evaluation, the AudioMOS Challenge introduced MOS prediction as a supervised learning task, encouraging the development of models that can approximate human judgments [14]. Building on this, the MusicEval benchmark expanded the evaluation framework by incorporating textual conditioning, requiring models to jointly assess both musical quality and textual alignment [6]. This shift reflects the growing importance of controllable music generation and the need for evaluation protocols that account for both audio fidelity and semantic correspondence to user intent. Recent research continues to explore hybrid approaches that combine objective and subjective metrics, as well as novel evaluation protocols that better reflect the creative and perceptual aspects of music. Nevertheless, developing reliable, scalable, and musically meaningful evaluation methods remains a central challenge for the field.

B. Contrastive Audio-Text Models

Recent progress in multimodal contrastive learning has led to models such as Contrastive Language-Audio Pretraining (CLAP), which aligns audio and text representations in a shared embedding space [14]. CLAP has demonstrated strong performance on a variety of audio-text retrieval tasks and has been widely adopted as a feature extractor for downstream applications, including MOS prediction for generative music [6], [15]. However, CLAP's training objective emphasizes coarse-grained audio-text matching and does not explicitly model detailed musical structure, leading to suboptimal sensitivity to perceptual aspects such as rhythm, melody, and timbre that are critical for expert human evaluation.

CLAP consists of two main components: an audio encoder and a text encoder. The audio encoder is based on a hierarchical token semantic audio transformer (HTS-AT) architecture, specifically employing a multi-layer Swin Transformer to process log-Mel spectrograms of audio signals [16]. The Mel-spectrogram input is divided into patches, which are then embedded and passed through cascaded Swin Transformer layers to extract hierarchical audio features. This design enables the model to capture both local and global time-frequency patterns in the audio [16], [17]. For the text modality, CLAP utilizes a RoBERTa-based encoder to extract semantic features from natural language descriptions [18]. Both the audio and text encoders project their respective features into a common latent space of identical dimension. During pretraining, CLAP uses a contrastive loss to maximize the similarity between paired audio and text samples while minimizing it for non-matching pairs, effectively aligning the modalities in the shared embedding space [16], [19].

By leveraging this dual-encoder architecture and contrastive training objective, CLAP is able to learn joint representations that bridge the gap between audio and language. Its large-scale pretraining leverages hundreds of thousands of audio-text pairs from diverse sources, enabling strong zero-shot and transfer capabilities across tasks such as text-to-audio retrieval, zero-shot audio classification, and supervised audio classification [18], [19]. Despite these strengths, the model's focus on global semantic alignment means it may overlook fine-grained musical details, which are crucial for perceptual quality assessment and expert-level music evaluation.

C. Self-Supervised Masked Audio Modeling

To overcome the limitations of contrastive objectives, self-supervised masked modeling approaches have gained increasing attention for learning rich acoustic representations [6], [9]. Inspired by masked language modeling in NLP, models such as M2D2 (Masked Modeling Duo) train encoders to reconstruct missing audio segments, forcing them to learn fine-grained temporal, spectral, and structural properties of audio signals [8].

Beyond M2D2, several other self-supervised masked modeling methods have been proposed in the audio domain. Wav2Vec 2.0 [20] is a prominent example, where a portion of the raw audio waveform is masked and the model is trained to predict the masked content based on the surrounding context. This approach enables the model to capture both local and global dependencies in the audio signal, leading to robust representations for downstream tasks such as speech

recognition and audio classification. HuBERT [21] extends this idea by first clustering acoustic features to create discrete targets and then training the model to predict these targets for masked segments. This method combines masked prediction with unsupervised clustering, further enhancing the model's ability to capture high-level semantic information from audio. Data2Vec [22] generalizes masked modeling by encouraging the model to predict contextualized latent representations of masked regions, rather than reconstructing the original input or predicting discrete labels. This allows the model to learn more abstract and task-agnostic representations. Additionally, methods such as BEATs [23] and MAE-AST [24] adapt masked autoencoding strategies from vision and language to the audio spectrogram domain, masking patches or frames and training the model to reconstruct the missing parts. These approaches have demonstrated strong performance on a variety of audio understanding tasks. Collectively, these self-supervised masked modeling techniques have advanced the field by enabling models to learn powerful and generalizable audio representations without requiring large amounts of labeled data. They are now foundational in state-of-the-art systems for speech, music, and general audio processing. Although masked modeling has shown promise in speech and general audio tasks, its application to music quality evaluation remains largely unexplored.

D. MOS Prediction Architectures

To leverage upstream features for MOS prediction, researchers have proposed a variety of neural architectures, from simple MLP regressors operating on frozen embeddings to fully end-to-end trainable models. WhisQw/OT introduces a sequence-level co-attention architecture between audio and text embeddings, combined with a Sinkhorn Optimal Transport regularization to enforce semantic alignment which resulting in significantly improved alignment performance on text-conditional music MOS prediction tasks [9]. However, the model still relies on frozen pretrained encoders for both audio (Whisper-Base) and text (Qwen-3), which limits its ability to adapt representations specifically for the MOS prediction task and to capture fine-grained perceptual cues unique to music.

In evaluating such models, the Mean Opinion Score (MOS) is widely used as the primary metric for subjective quality assessment in audio and music generation tasks. MOS reflects the average rating given by human listeners on a Likert scale (typically 1 to 5), capturing perceived quality in a way that closely aligns with human judgment. In the context of music generation, MOS is collected for two key dimensions: overall musical impression and textual alignment, providing a comprehensive view of both the intrinsic musical quality and the relevance to the given text prompt.

To ensure a thorough assessment of model performance, evaluations are conducted at both the utterance and system levels. Utterance-level evaluation refers to computing metrics at the level of individual music clips, where the model predicts a quality score for each generated sample, and these predictions are directly compared to the corresponding human ratings. This level of granularity is crucial for assessing how well a model can evaluate the quality of each unique piece of music, capturing sample-specific nuances. In contrast, system-level evaluation aggregates predictions and human ratings across all samples generated by a particular model or system. This approach provides a higher-level view of a model's ability to rank or assess the overall performance of different generative systems, which is important for benchmarking and comparing systems in a fair and consistent manner. Together, these evaluation strategies offer a robust framework for measuring the effectiveness of MOS prediction architectures in the context of text-to-music generation [6], [9].

The quality of MOS prediction models is further assessed using a set of objective metrics that provide a comprehensive evaluation of model performance. These metrics capture both the accuracy of the predicted scores and the consistency of ranking across samples and systems:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual MOS values, quantifying the accuracy of the predictions.
- Linear Correlation Coefficient (LCC): Assesses the linear relationship between predicted and ground-truth MOS scores, indicating how well the model's predictions follow the actual ratings.
- Spearman Rank Correlation Coefficient (SRCC): Evaluates the monotonic relationship between predicted and actual rankings, reflecting the model's ability to preserve the order of quality across samples or systems.
- Kendall's Tau (KTAU): Another rank-based metric that measures the strength of association between predicted and ground-truth rankings, providing additional insight into ranking consistency.

Among these, **MSE and LCC are generally the most important metrics when the goal is to predict the exact quality scores**, as they directly measure the accuracy and trend alignment between predicted and actual MOS values. **SRCC and KTAU become especially important when the primary concern is the correct ranking of samples or systems**, such as in benchmarking or recommendation settings. By employing all four metrics, researchers can systematically compare different MOS prediction models, capturing both the absolute accuracy of the predicted scores and the fidelity of ranking at various levels of granularity. In this work, we advance the state-of-the-art by combining the representational strength of M2D2 with a dual-branch convolutional architecture, enabling richer feature extraction and joint modeling of both audio and text for improved MOS prediction.

III. METHODOLOGY

We aim to develop an automated system capable of predicting the human-assessed quality of AI-generated music. Our evaluation is grounded in two expert-defined criteria:

1) **Overall Musical Impression:** This criterion refers to the general aesthetic and emotional impact of the music. Specifically, the score takes into account factors such as the authenticity of the music and the quality of the melody, rhythm, and chords. A low score indicates that the sample lacks musicality and is of very poor quality,

or exhibits noticeable machine-generated artifacts. In contrast, a high score suggests that the sample is of excellent overall quality, characterized by a clear rhythm, a pleasant melody, and a coherent chord progression, making it difficult to distinguish whether it was composed by a human or generated by a system.

2) **Textual Alignment:** This criterion assesses how well the generated music matches or complements a given text prompt. It examines whether the mood, tempo, and style of the music appropriately reflect the content, emotion, and intent of the text. The textual alignment score evaluates how well the audio sample corresponds to the given text description, reflecting the system's ability to adhere to the prompt. A low score indicates little or no relevance between the generated music sample and the text description, while a high score suggests a strong alignment between the two.

By predicting these two aspects, our goal is to create a model that closely mirrors human judgment in evaluating both the intrinsic musical quality and the appropriateness of the music in relation to its intended textual context.

A. Dataset

We assess our approach on Track 1 of the AudioMOS 2025 Challenge, utilizing the expert-annotated dataset released by the MusicEval benchmark [6]. The MusicEval dataset comprises 2,748 generated music clips produced by 31 different text-to-music models, covering a broad spectrum of generative systems. The dataset features 384 unique text prompts, including 80 manually crafted prompts, 20 from the MusicCaps dataset, and 284 extracted from system demo pages. These prompts span a range of musical aspects, such as emotion, structure, rhythm, theme, and instrumentation. They focus on the pop and classical genres to leverage the evaluators' expertise. For evaluation, 14 raters (2 professional teachers and 12 experienced students from conservatories) assessed each music clip. All ratings were collected using a 5-point Likert scale, with each audio sample evaluated by five different raters. To ensure reliability, the evaluation protocol incorporated quality control measures, including the insertion of real human-created music clips and duplicate samples to identify and filter out inconsistent ratings. In total, the benchmark includes 13,740 high-quality ratings, providing a robust foundation for assessing the performance of text-to-music generation systems.

B. Model Architecture

Our model, ConvM2D2, employs a dual-branch architecture that independently encodes audio and text inputs before fusing their representations for joint regression tasks. We utilize a pre-trained second-generation Masked Modeling Duo (M2D2-CLAP) encoder to extract D-dimensional embeddings from both the audio signal and its corresponding text prompt. These embeddings are projected into a shared latent space to facilitate semantic alignment between text and the acoustic features of the audio.

The audio branch includes a feature extractor composed of four sequential convolutional blocks. Each block consists of a $1D$ convolution layer (kernel size = 3, stride = 1, padding = 1), followed by a ReLU activation and a max-pooling operation (kernel size = 2). The number of output channels in the convolutional layers increases as follows: 64, 128, 256, and C respectively. This results in a feature map of shape $(\text{batch\_size}, C, T')$, where T' is the temporally downsampled sequence length. A global average pooling operation is applied across the temporal dimension to yield a final C-dimensional audio feature vector.

For the alignment pathway, the audio feature is concatenated with the corresponding D-dimensional text embedding to produce a joint $2D$-dimensional representation. Two separate 3-layer multilayer perceptrons (MLPs) are then used to generate scalar predictions:

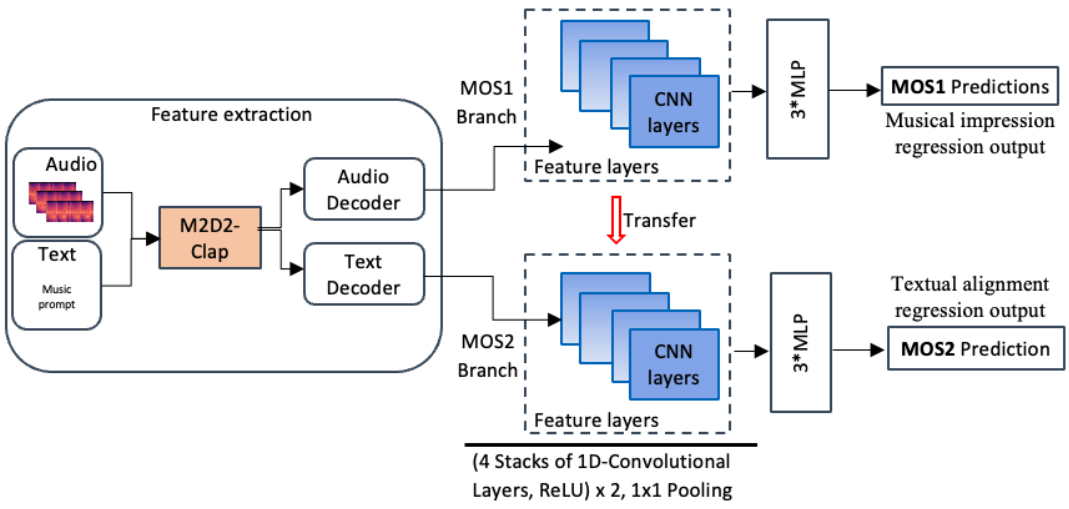

Fig. 2. Overview of the proposed ConvM2D2 architecture. The input audio is transformed into a log-mel spectrogram and paired with its corresponding text prompt. Both modalities are encoded using pretrained M2D2 models. The resulting embeddings are fused via sequence-level coattention, followed by two lightweight MLP heads that independently predict the overall musical quality (MOS1) and textual alignment score (MOS2).

- Overall Quality Branch: An MLP that takes the C-dimensional audio feature as input and outputs a single score.
- Textual Alignment Branch: An MLP that takes the $2D$-dimensional audio-text feature and outputs a single alignment score.

C. Model Formulation

Let $x_a \in \mathbb{R}^T$ denote the raw audio input, and $x_t \in \mathbb{R}^P$ denote the corresponding text prompt. Both inputs are first passed through pretrained encoders to obtain their global embeddings:

$$z_a = E_{\text{audio}}(x_a) \in \mathbb{R}^D,$$
$$z_t = E_{\text{text}}(x_t) \in \mathbb{R}^D.$$

To enhance local feature extraction, the raw audio input x_a is further processed through a convolutional feature extractor:

$$h_{\text{conv}} = \text{ConvBlock}(x_a),$$

where $\text{ConvBlock}(\cdot)$ consists of a sequence of one-dimensional convolution, activation, and pooling operations. After global average pooling along the temporal axis, we obtain:

$$f_a = \text{GlobalPool}(h_{\text{conv}}) \in \mathbb{R}^C.$$

For the textual alignment branch, the audio and text features are concatenated to form a joint representation:

$$f_{\text{align}} = \begin{bmatrix} f_a \\ z_t \end{bmatrix} \in \mathbb{R}^{C+D}.$$

Two independent multilayer perceptrons (MLPs) then produce the quality predictions:

$$\hat{y}_{\text{overall}} = \text{MLP}_{\text{overall}}(f_a),$$
$$\hat{y}_{\text{align}} = \text{MLP}_{\text{align}}(f_{\text{align}}).$$

D. Training Procedure

The network is optimized in an end-to-end manner using mini-batch stochastic gradient descent (SGD). A loss of the mean absolute error (L1) is employed, applied independently to each of the two prediction branches of the model.

At each training step, the overall Mean Opinion Score (MOS) prediction is supervised using:

$$L_{\text{overall}} = \left|\hat{y}_{\text{overall}} - y_{\text{overall}}\right|,$$

and the alignment (or coherence) prediction is supervised using:

$$L_{\text{align}} = \left|\hat{y}_{\text{align}} - y_{\text{align}}\right|.$$

The total training loss is computed as the average of the two:

$$L_{\text{total}} = \frac{1}{2}\left(L_{\text{overall}} + L_{\text{align}}\right).$$

We use SGD with a momentum coefficient of 0.9 and a learning rate of 5×10^{-4}. No weight decay is applied, as preliminary experiments indicated that regularization was not necessary to prevent overfitting in our setting. This may be due to the size of our dataset and regularization. We use a mini-batch size of 32 for training and 8 for validation, which strikes a practical balance between stable model convergence and GPU memory constraints imposed by the large size of full audio and text embeddings, consistent with findings that smaller batch sizes often improve generalization and training stability [25]. The model is trained for up to 1000 epochs with early stopping, where training terminates if the validation loss does not improve for 20 consecutive epochs, and the checkpoint with the lowest validation loss is selected as the best-performing model. To enhance robustness, we employ 10-fold cross-validation on the training set with data shuffled using a fixed random seed; each fold is used once for validation while the remaining nine are used for training. Final validation performance is reported on a separate held-out development set. All training and evaluation are performed on a single GPU-equipped server 1 GPU (NVIDIA RTX A5000), with CUDA support. Evaluation Metrics include Mean Squared Error (MSE), Pearson (LCC), Spearman (SRCC), and Kendall Tau (KTAU) correlations. However, training progress is monitored using the average $L1$ loss on the validation set, as we observed no significant difference in performance when compared to other loss metrics. MAE also provides a straightforward and interpretable measure of prediction error, making it a practical choice for this task.

IV. Experimental Results

We evaluate the performance of our proposed ConvM2D2 model on the MusicEval Track-1 validation set. The evaluation covers both the Overall Musical Quality and Textual Alignment tasks, assessed at both utterance-level and system-level. Performance is measured using Mean Squared Error (MSE), Linear Correlation Coefficient (LCC), Spearman Rank Correlation Coefficient (SRCC), and Kendall's Tau (KTAU) metrics. Table 1 presents the comprehensive results across four models: (1) the baseline model provided by the MusicEval Track-1 challenge organizers [6]; (2) Conv-Augmented Baseline (CAB), which extends the baseline model with an additional convolutional block to enhance local feature extraction [9]; (3) WhisQw/OT, a state-of-the-art model for music quality assessment [9]; and (4) our proposed ConvM2D2 model.

A. Performance Analysis

Across all evaluation metrics, ConvM2D2 consistently outperforms the baselines, achieving substantial improvements. In particular, for the SRCC metric on overall musical impression, ConvM2D2 achieves a score that is approximately 35% higher than the baseline.

To investigate the impact of local feature extraction, we augmented the baseline model by introducing a convolutional block, resulting in the Conv-Augmented Baseline (CAB) model. The convolutional layer enhances the model's ability to capture local acoustic patterns such as noise bursts, clipping, and transient distortions often overlooked by global representations. The results demonstrate that capturing local details, such as brief distortions or irregularities in the audio, provides

TABLE I
COMPREHENSIVE EVALUATION RESULTS ON MUSICEVAL TRACK-1 TESTING DATASET

	OVERALL_MUSICAL_QUALITY								**TEXTUALL_ALIGNMENT**							
	Utterance-level				System-level				Utterance-level				System-level			
	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑
Baseline	0.698	0.603	0.615	0.452	0.378	0.821	0.818	0.623	0.585	0.514	0.521	0.375	0.199	0.744	0.724	0.532
WhisQw/OT	0.349	0.744	0.738	0.555	0.144	0.858	0.869	0.706	0.464	0.570	0.572	0.415	0.147	0.790	0.807	0.615
CAB	0.328	0.791	0.791	0.604	0.103	0.937	0.939	0.813	0.424	0.648	0.628	0.464	0.057	0.894	0.874	0.700
ConvM2D2	**0.265**	**0.825**	**0.831**	**0.649**	**0.044**	**0.963**	**0.965**	**0.848**	**0.433**	**0.659**	**0.655**	**0.486**	**0.070**	**0.926**	**0.904**	**0.767**

valuable information that global pretrained models alone may overlook. By integrating these local features with the broader contextual understanding from global embeddings, the model achieves a more comprehensive and accurate assessment of musical quality.

Compared to WhisQw/OT, which also uses powerful pretrained models for both audio and text, ConvM2D2 delivers a 9.5% improvement in performance across all evaluation metrics. Although WhisQw/OT benefits from extensive pretrained knowledge, it mainly focuses on global information and may miss subtle musical issues that affect perceived quality. ConvM2D2 addresses this by combining global context from pretrained models with additional local features extracted by the convolutional layer. This allows the model to detect small distortions, timing issues, and other detailed quality problems that might otherwise be overlooked. As a result, ConvM2D2 provides a more balanced and accurate assessment of musical quality.

By providing a robust, self-supervised framework for evaluating generative music systems, ConvM2D2 empowers researchers, developers, and artists to objectively assess musical outputs at scale without the need for costly human annotation. This enables more transparent benchmarking and fairer comparisons across models, accelerating progress in music AI research and commercial applications. Beyond objective assessment, ConvM2D2 democratizes access to advanced music evaluation, empowering a wider range of creators, educators, and researchers to experiment with and refine generative music systems. Its integration into creative workflows supports rapid prototyping and new forms of human-AI collaboration. By establishing transparent, reproducible benchmarks, ConvM2D2 also fosters responsible innovation and fair competition in both academia and industry. Furthermore, its use can help address ethical, economic, and cultural challenges posed by the proliferation of AI-generated music, ensuring that technological progress aligns with the broader interests of the music community.

V. CONCLUSION

In this work, we present ConvM2D2, a dual-branch neural architecture for automatic prediction of human-judged quality of AI-generated music, jointly considering both overall musical impression and textual alignment between audio and text prompts. Leveraging a pre-trained M2D2-CLAP encoder to extract modality-specific embeddings, our model effectively integrates audio and text information through shared latent projections and specialized regression heads for each evaluation criterion.

We conduct comprehensive experiments on the AudioMOS Challenge dataset, comparing ConvM2D2 against both the official MusicEval Baseline, CAP and the WhisQw/OT system across a wide range of metrics, including MSE, LCC, SRCC, and KTAU, at both utterance-level and system-level evaluations. Results demonstrate that ConvM2D2 achieves consistent and substantial improvements across all metrics. Notably, we observe reductions in MSE of up to 88% and significant gains in correlation metrics, highlighting ConvM2D2's capacity to more accurately capture both absolute quality judgments and relative ranking consistency in subjective music evaluation.

The dual-branch architecture, together with effective fusion of audio and text embeddings, enables the model to better align its predictions with human expert assessments. These improvements suggest that ConvM2D2 can serve as a reliable automated proxy for subjective music evaluation, which is critical for scalable benchmarking of AI-generated music models. This advancement also enables large-scale, consistent, and cost-efficient evaluation of generative music systems, reducing dependence on time-consuming and resource-intensive expert listening tests.

In the future, we plan to extend the proposed framework to explore advanced alignment objectives (contrastive learning, cross-modal consistency losses), and evaluate its generalizability across larger, more diverse generative music datasets. We also aim to investigate real-time deployment possibilities for continuous monitoring of generative music systems.

REFERENCES

[1] P. Dhariwal, H. Jun, J. W. Payne, C. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *OpenAI Blog*, vol. 1, 2020. [Online]. Available: https://openai.com/blog/jukebox

[2] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 704–47 720, 2023.

[3] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.

[4] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The voicemos challenge 2024: Beyond speech quality prediction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 803–810.

[5] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," *arXiv preprint arXiv:2208.12208*, 2022.

[6] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin, "Musiceval: A generative music corpus with expert ratings for automatic text-to-music evaluation," *arXiv preprint arXiv:2501.10811*, 2025. [Online]. Available: https://arxiv.org/abs/2501.10811

[7] A. Lerch, C. Arthur, N. Bryan-Kinns, C. Ford, Q. Sun, and A. Vinay, "Survey on the evaluation of generative models in music," *arXiv preprint arXiv:2506.05104*, 2025.

[8] D. Niizumi, D. Takeuchi, M. Yasuda, B. T. Nguyen, Y. Ohishi, and N. Harada, "M2d2: Exploring general-purpose audio-language representations beyond clap," *arXiv preprint arXiv:2503.22104*, 2025. [Online]. Available: https://arxiv.org/abs/2503.22104

[9] J. I. Emon, K. T. Alam, and M. A. Salek, "Whisq: Cross-modal representation learning for text-to-music mos prediction," *arXiv preprint*, 2025.

[10] Z. Xiong, W. Wang, J. Yu, Y. Lin, and Z. Wang, "A comprehensive survey for evaluation methodologies of AI-generated music," *arXiv preprint arXiv:2308.13736*, 2023.

[11] A. Hilmkil, C. Thomé, and A. Arpteg, "Perceiving music quality with GANs," *arXiv preprint arXiv:2006.06287*, 2020.

[12] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[13] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting fréchet audio distance for generative music evaluation," *arXiv preprint arXiv:2311.01616*, 2023.

[14] S. Bae, "Audiomos challenge 2025," https://www.linkedin.com/posts/soohyun_voicemos-challenge-audiomos-challenge-2025-activity-7313314395181195264-_ocS, 2025, accessed: 2025-06-23.

[15] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022. [Online]. Available: https://arxiv.org/abs/2206.04769

[16] Z. Wu *et al.*, "Clapsep: Leveraging contrastive pre-trained model for multi-source audio separation," *arXiv preprint arXiv:2402.17455*, 2024. [Online]. Available: https://arxiv.org/abs/2402.17455

[17] Y. Wu, T. Zhang, and K. Chen, "Text-to-audio retrieval via large-scale contrastive training," *IEEE AASP Challenge Detection Classification Acoust. Scenes Events, Nancy, France*, 2022.

[18] UnrealSpeech, "Introduction to clap: Unveiling the symphony of language and sound," 2024, https://blog.unrealspeech.com/introduction-to-clap-unveiling-the-symphony-of-language-and-sound/.

[19] LAION-AI, "Laion-ai/clap: Contrastive language-audio pretraining," 2023, https://github.com/LAION-AI/CLAP.

[20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3451–3460. [Online]. Available: https://arxiv.org/abs/2106.07447

[22] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning (ICML)*, 2022. [Online]. Available: https://arxiv.org/abs/2202.03555

[23] S. Chen, Y. Wang, J. Li, C. Xuankai, C. Li, Y. Wu, Z. Yao, X. Wang, S. Liu, J. Shi *et al.*, "Beats: Audio pre-training with acoustic tokenizers," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2212.09058

[24] S. Baade, Y. Wang, J. Li, Y. Wu, C. Xuankai, X. Wang, S. Liu, J. Shi *et al.*, "Mae-ast: Masked autoencoding audio spectrogram transformer," in *arXiv preprint arXiv:2306.02948*, 2023. [Online]. Available: https://arxiv.org/abs/2306.02948

[25] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.