Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AImoclips: Benchmarking Emotion Conveyance in Text-to-Music Generation Using a Dimensional Valence-Arousal Framework

Gyehun Go
Korea Advanced Institute of Science and Technology
Satbyul Han
Seoul National University
Ahyeon Choi
University of Iowa
Eunjin Choi
Korea Advanced Institute of Science and Technology
Juhan Nam
Korea Advanced Institute of Science and Technology
Jeong Mi Park
mensa04@snu.ac.kr

Seoul National University

Article

Keywords:

Posted Date: May 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9027741/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Almoclips: Benchmarking Emotion Conveyance in Text-to-Music Generation Using a Dimensional Valence-Arousal Framework

Gyehun Go[1], Satbyul Han[2], Ahyeon Choi[3], Eunjin Choi[1], Juhan Nam[1], and Jeong Mi Park[2*]

[1]Korea Advanced Institute of Science and Technology, Graduate School of Culture Technology, Daejeon, 34141, South Korea
[2]Seoul National University, Department of Music, Seoul, 08826, South Korea
[3]University of Iowa, Department of Communication Sciences and Disorders, Iowa City, IA 52240, USA
[*]mensa04@snu.ac.kr

ABSTRACT

Recent advances in text-to-music (TTM) generation have enabled controllable and expressive music creation using natural language prompts, yet the extent to which these systems faithfully convey intended emotions remains largely underexplored. In this study, we introduce Almoclips, a benchmark designed to evaluate emotion conveyance to human listeners in TTM systems using a dimensional valence-arousal framework. We constructed the dataset with 991 instrumental music clips from six TTM systems prompted with 12 emotion words spanning four valence-arousal quadrants. A total of 111 participants provided 6,162 valence and arousal ratings using a 9-point scale. The results revealed that all systems perform above chance in conveying quadrant-level emotion intent, yet overall accuracy remains limited. Notably, substantial model-dependent biases were present. Commercial systems tended to exhibit positive valence deviations, whereas open-source models more often produced outputs with a negative valence shift, with diverging arousal shifts. Furthermore, overall audio quality (Fréchet Audio Distance (FAD)) correlates with both valence and arousal, while text-audio alignment (Contrastive Language-Audio Pretraining (CLAP) score) primarily relates to valence. These findings highlight underlying challenges in conveying linguistic emotional semantics precisely through musically conveyed affect for future research on controllable and perceptually consistent music generation.

Introduction

Recent advances in generative AI have led to rapid progress in creative domains, including music. Early work focused on unconditional generation, producing music without explicit user control[1–3]. While these models first showed the feasibility of generating high-quality music, they offered limited mechanisms for specifying user intent. Subsequent work introduced conditional generation, allowing control over musical attributes such as style, structure, or semantic descriptors[4–6].

More recently, large-scale foundation models have demonstrated that conditioning music generation on text is a powerful and flexible paradigm[7], leading to the development of a novel approach called text-to-music (TTM) generation[8–12]. Textual prompts were widely adopted due to their familiarity and expressive capacity to describe a diverse array of subjective emotions and musical concepts. In particular, commercial TTM systems such as Suno[13] and Udio[14] have greatly enhanced the quality and accessibility of music generation so that non-expert users can simply generate high-quality music through simple free-form textual input. Since then, the use of TTM systems has rapidly increased in content creation; it has become crucial to understand how effectively they convey textual user intent into perceptually meaningful musical output and satisfy users' demand for generated music.

Existing works to evaluate TTM systems have largely focused on human preference and text–audio alignment. Grötschla et al.[15] conducted a large-scale benchmark of music generation models and examined which settings of commonly used automatic metrics for music generation evaluation can better align with human preference judgments. Other studies have suggested using learned evaluation models[16] or metrics[17] that better align with the way humans perceive music, including datasets with expert ratings representing user preferences. These efforts reflect a common assumption: evaluation metrics should prioritize perceptual quality and alignment with the text prompt.

While these approaches suggest ways to improve the evaluation of TTM systems, they often overlook emotional aspects of generated music. Since emotional impression plays a central role in how listeners experience and remember music, overlooking emotion conveyance may lead to mismatches between user expectations and generated outputs, limiting the applicability of TTM systems in real-world contexts such as affective music recommendation, adaptive soundtracks, or interactive media. Music preference, audio quality, and emotional communication are not equivalent; a piece of music may be technically well

produced and aligned with a text prompt while failing to convey the intended emotional properties, or conversely, may evoke an intended, strong emotional response despite modest perceived quality.

To assess emotional responses to music, prior research has employed the frameworks of discrete or dimensional models. Discrete emotion models represent emotions as categorical states (e.g., happy, sad, angry) and have been widely used in music emotion recognition and tagging datasets[18–21]. In contrast, dimensional models depict emotions as points on a space constructed by continuous axes[22]. One commonly used model, the circumplex model of affect[23,24], maps emotion onto 2 dimensions: valence and arousal. The valence–arousal (VA) framework has been widely adopted in music emotion research due to its simplicity, interpretability, and ability to capture gradual affective changes.

A substantial body of work has investigated emotion-aware music generation utilizing both discrete and continuous representations[25–30]. These studies demonstrate that conditioning generative models on emotional attributes is feasible, thereby facilitating the control of conveyed emotional characteristics. However, most prior work has focused on emotion-aware music generation rather than perceptual evaluation, and relatively few studies have examined how well generated music actually conveys intended emotions to human listeners in a systematic and quantitative manner.

A recent study[31] represents one of the few attempts to directly evaluate the emotion conveyance ability of music generation systems. Their work examined "emotion conversion accuracy" by presenting listeners with multiple-choice questions with 4 options that correspond to four quadrants of the valence–arousal plane. While this approach showed that current systems can convey coarse emotional categories above chance, the evaluation was limited to the assessment data with forced-choice categorical annotations. Such an evaluation format prevented more fine-grained analyses of conveyed emotional attributes, such as distributions of conveyed emotions on affective space and the direction of each system's bias.

In this study, we address these limitations by collecting dimensional ratings on conveyed emotion from text-to-music systems. Using the valence–arousal framework, we collected two-dimensional scores from human listeners using a 9-point Self-Assessment Manikin (SAM) scale. We introduce AImoclips, a new open-source dataset that consists of 991 music clips generated by six contemporary TTM systems, along with the valence-arousal ratings on them from a total of 111 participants (detailed statistics of valence and arousal ratings are available in Supplementary Table S1 and S5). Each system was prompted with 12 emotion words—three per valence-arousal quadrant—to generate the dataset from a balanced set of emotion intents. Statistical analysis reveals not only quadrant-level emotion conveyance accuracy but also distributional properties and system-specific biases that cannot be captured by categorical evaluation alone. We also examine how perceived emotion is associated with objective evaluation metrics, like Fréchet Audio Distance (FAD)[32] and Contrastive Language-Audio Pretraining (CLAP)-based semantic alignment scores[33], to examine which factors are associated with emotion communication in TTM systems.

This study examines how effectively TTM systems convey intended emotional content to human listeners through musical output by addressing the following research questions: (1) How effectively do TTM systems convey intended emotional content to human listeners? (2) How does emotion conveyance differ across TTM systems and across regions of the valence–arousal space? (3) What are key factors that are related to the emotion conveyance of TTM systems? By answering these questions, we aim to provide a systematic benchmark of emotion conveyance in TTM systems and to clarify the current state, highlighting capabilities and limitations of current approaches to emotion-aware music generation.

Results

Participants and Scale of Dataset

A total of 111 participants rated the generated music clips. After excluding clips with three or fewer raters, this yielded 6,162 total ratings. The detailed distribution of the number of ratings per clip is shown as Supplementary Figure S1 online. Six ratings from one participant were removed due to a technical issue on the survey website that prevented them from listening to the final six samples. The participant group consisted of 64 females and 47 males. Most participants (89 out of 111) were young adults aged 18 to 29. The remaining participants included middle-aged individuals, specifically six aged 45–49 and five aged 50 or older.

Overall Emotion Conveyance of TTM Systems

To evaluate the emotional conveyance of TTM systems, we first employed a categorical approach by calculating the mean valence and arousal ratings for each clip. We divided these ratings into four groups (Q1-Q4 defined in Table 1), with a neutral threshold of 5.0 for both dimensions. We then calculated the degree of alignment between the intended emotion quadrant and the rated emotion quadrant. If a clip's mean rating for either valence or arousal was exactly 5.0, it was categorized as 'Neutral.' All neutral ratings were considered a mismatch regardless of the intended emotion quadrant. Table 1 presents the calculated precision, recall, and F1 scores for each quadrant. The overall accuracy of all the systems was 0.4097, and the weighted F1 score was 0.4279. Although over 50% of the produced clips did not effectively communicate the intended emotional quadrant to listeners, the accuracy significantly exceeded chance ($z = 10.22, p < .001$). An examination of the individual quadrants

Table 1. Categorical performance metrics for emotion conveyance for each valence-arousal quadrant. Precision, recall, and F1 scores are calculated based on the alignment between intended emotion quadrants and rated emotion quadrants for all 6,162 ratings. *n* represents the number of clips generated from an emotion word in each quadrant.

Quadrant	Valence	Arousal	Precision	Recall	F1	*n*
Q1	High	High	0.597	0.524	0.558	246
Q2	Low	High	0.549	0.294	0.383	248
Q3	Low	Low	0.338	0.472	0.394	250
Q4	High	Low	0.411	0.348	0.377	247

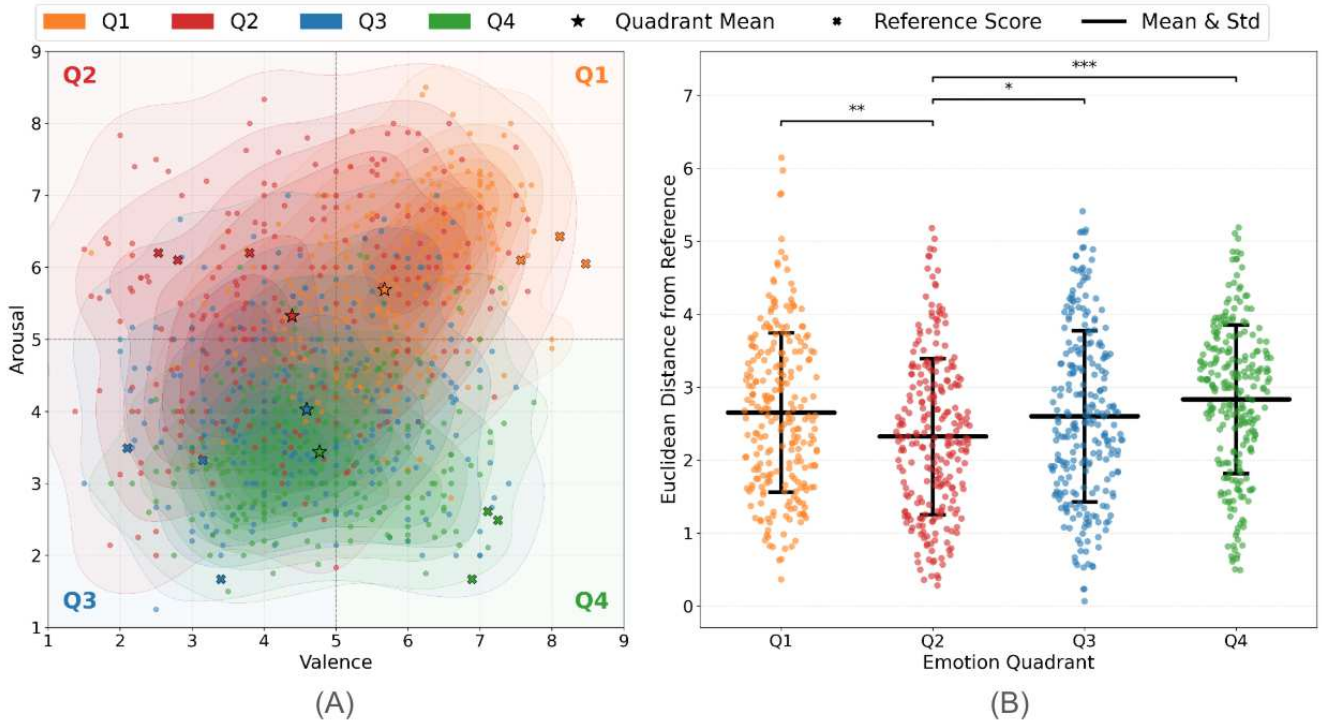


Figure 1. (A) Valence-arousal rating distributions of all clips in AImoclips, with KDE contours for each set of clips by their quadrant of intended emotion. Stars represent mean ratings for each quadrant, and 'X' marks denote the reference score for the emotion intents. (B) Distribution plot of Euclidean distance between the rated scores and the reference scores for each quadrant. Asterisks ($*p < .05$, $**p < .01$, $***p < .001$) denote statistical significance levels of difference between quadrants.

indicated that emotions in Q1 (high valence, high arousal) were expressed with the highest accuracy. Despite the F1 scores being relatively comparable in the other quadrants, we observed that Q3 (low valence, low arousal) exhibited substantially lower precision than Q2 (low valence, high arousal) and Q4 (high valence, low arousal).

We also conducted a quantitative analysis by calculating the Euclidean distance between the mean clip ratings and the reference scores of emotion words used from the extension[34] of Affective Norms for English Words (ANEW)[35]. A one-way ANOVA showed that the mean distances between the quadrants had significant differences ($F(3,987) = 9.283, p < .001, \eta^2 = .027$). Levene's test confirmed equal variances across quadrants ($W = 2.43, p = .064$). Although the Shapiro-Wilk test indicated non-normality of residuals ($W = 0.993, p < .05$), inspection of residual distributions indicated no substantial departure from normality, and the large sample size ($n = 991$) supports the robustness of the ANOVA results. After establishing overall differences among the four quadrants using a one-way ANOVA, we further examined whether these differences were driven by valence, arousal, or their interaction by conducting a two-way ANOVA. The subsequent two-way ANOVA revealed that both valence ($F(1,987) = 16.37, p < .001, \eta_p^2 = .016$) and arousal levels ($F(1,987) = 11.03, p < .001, \eta_p^2 = .011$) had significant effects, respectively, but there was no significant interaction ($F(1,987) = 0.47, p = .491, \eta_p^2 < .001$). While Q2 demonstrates the least mean distance to the reference score, Q1, which exhibited the highest categorical accuracy, possesses a greater Euclidean distance relative to other quadrants (Figure 1(B), see Supplementary Table S2 for detailed statistics).

Overall, the clips were generally rated more neutrally than the original emotion words used in the text, exhibiting a centralizing tendency. According to Figure 1(A) and Figure 3, paired t-tests showed that the mean distance from the center of the VA plane (5,5) to clip rating scores was significantly shorter than to the reference scores of emotion intents for all TTM systems, both within each subset of clips from each TTM system and the overall dataset ($p < .001$). A Shapiro-Wilk test on the overall paired differences indicated non-normality ($W = 0.995, p < .05$); however, the large sample size per system ($n \geq 161$) ensures robustness of the t-test. Detailed statistics about distances from the center can be found in Supplementary Table S4.

Table 2. Comparative performance of TTM systems in emotion conveyance. The total accuracy of quadrant matching, quadrant-specific F1 scores, and weighted F1 scores are provided for each TTM system. The highest values in each column are marked **bold**, and the second highest values are underlined.

TTM System	Accuracy	Weighted F1	F1 (Q1)	F1 (Q2)	F1 (Q3)	F1 (Q4)
AudioLDM 2	0.335	0.323	0.509	0.167	<u>0.443</u>	0.167
MusicGen	0.363	0.375	0.419	0.452	0.369	0.264
Mustango	<u>0.448</u>	<u>0.425</u>	0.462	<u>0.563</u>	**0.565**	0.103
Stable Audio Open	0.407	**0.434**	0.507	**0.571**	0.337	0.320
Suno	**0.463**	0.412	<u>0.635</u>	0.204	0.214	**0.604**
Udio	0.440	0.365	**0.656**	0.047	0.197	<u>0.565</u>

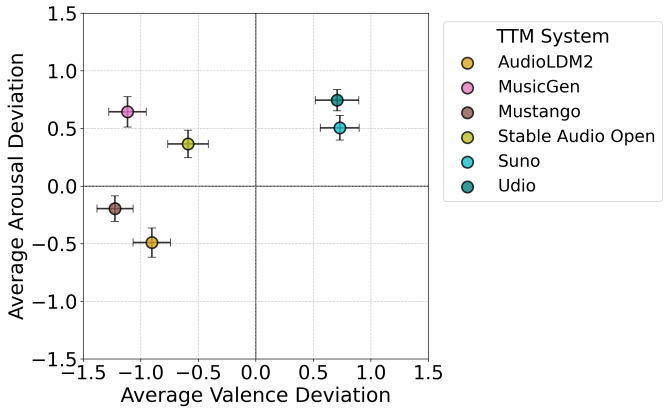


Figure 2. The scatter plot showing mean valence and arousal deviations from the reference score by TTM system. Horizontal and vertical error bars represent the standard errors for valence and arousal, respectively.

Difference in Emotion Conveyance Among TTM Systems

We applied the categorical analysis to each TTM system to compare their individual emotion conveyance capabilities, as summarized in Table 2 (confusion matrices for each TTM system can be found as Supplementary Figure S2). All systems performed above the random chance threshold of 25%, though none achieved an accuracy exceeding 50%. Suno demonstrated the highest accuracy (0.463), followed by Mustango (0.448). Distinct performance tendencies emerged between commercial and open-source systems. Commercial systems, including Suno and Udio, demonstrated proficiency, especially in expressing Q1 (high valence, high arousal) emotions. Conversely, Mustango demonstrated a greater capacity to express low-valence emotions, whereas Stable Audio Open excelled in conveying high-arousal emotions. AudioLDM2 and MusicGen generally exhibited lower accuracy across all quadrants compared to the other systems.

The varying degrees of emotional conveyance across the TTM systems were also revealed in a numerical way by the mean deviation of valence-arousal ratings from the reference scores (Figure 2). The most notable finding was that the generations from Suno and Udio produced higher valence ratings than the reference, whereas the open-source systems often produced music perceived as less pleasant than intended, resulting in lower valence ratings in comparison to the reference. This data establishes a clear distinction between open-source and commercial systems. Arousal deviations were less pronounced yet still exhibited two different tendencies of TTM systems: AudioLDM 2 and Mustango exhibited negative arousal deviations, whereas the others displayed positive deviations.

A two-way ANOVA with interaction analyzing variations in valence and arousal, employing the TTM system and emotional intent as factors, revealed significant interaction effects for both valence ($F(55,919) = 3.93, p < .001, \eta_p^2 = .190$)

and arousal ($F(55,919) = 6.96, p < .001, \eta_p^2 = .294$), indicating that TTM systems exhibit emotion-specific strengths and weaknesses in conveying the intended affective content. The main TTM system effect was significant for both valence ($F(5,919) = 175.45, p < .001, \eta_p^2 = .488$) and arousal ($F(5,919) = 52.72, p < .001, \eta_p^2 = .223$). Residual diagnostics indicated non-normality (Shapiro–Wilk: valence $W = 0.995, p < .05$; arousal $W = 0.995, p < .05$) and heteroscedasticity across TTM system groups (Levene's: valence $W = 4.23, p < .05S$; arousal $W = 6.05, p < .05$). However, given the nearly equal group sizes ($164 \leq n \leq 167$) and large per-group sample sizes ($n \geq 164$), the ANOVA F-tests are considered robust to these deviations under a balanced design. Pairwise comparisons using the Tukey-Kramer method confirmed these differences. In terms of valence, both Suno and Udio, those with positive valence deviations, exhibited a significant difference from Stable Audio Open ($p < .001$), indicating the least negative deviation. Mustango, the system with the least negative arousal deviation, also exhibited a statistically significant difference in arousal compared to those with positive deviation ($p = .007$ between Mustango and Stable Audio Open, and $p < .001$ between Mustango and Suno). In contrast, no pair of models within the same sign of valence and arousal deviation exhibited a meaningful difference in valence-arousal deviation.

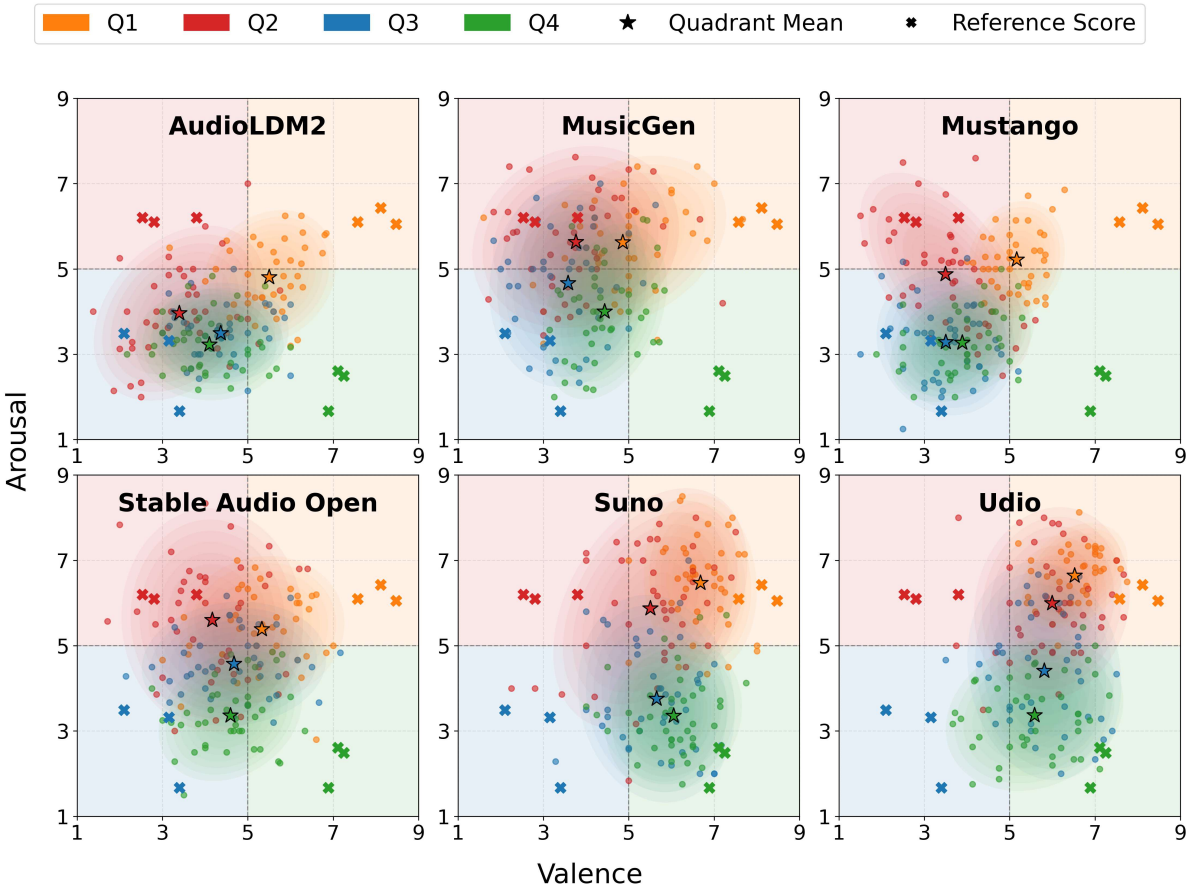


Figure 3. Valence-arousal rating distributions of all clips for individual TTM systems. Stars represent mean ratings for each quadrant; 'X' marks denote the reference score for the emotion intents; ellipses indicate the 95% confidence regions for score distribution of each quadrant.

Figure 3 visualizes the distribution of valence-arousal ratings for the set of clips produced by each TTM system (see Supplementary Figure S3 and Supplementary Table S3 for pairwise comparison results and detailed statistics, respectively). The ratings for clips generated by Suno and Udio exhibit a bias toward higher valence, suggesting that these models predominantly create music that expresses more pleasant emotions. Conversely, the rating distributions for open-source models generally favor lower valence, with AudioLDM2 and Mustango also displaying a tendency toward lower arousal. Another notable observation is that Suno and Udio exhibit vertically broader distributions, indicating that they produce music clips with a wide spectrum of arousal.

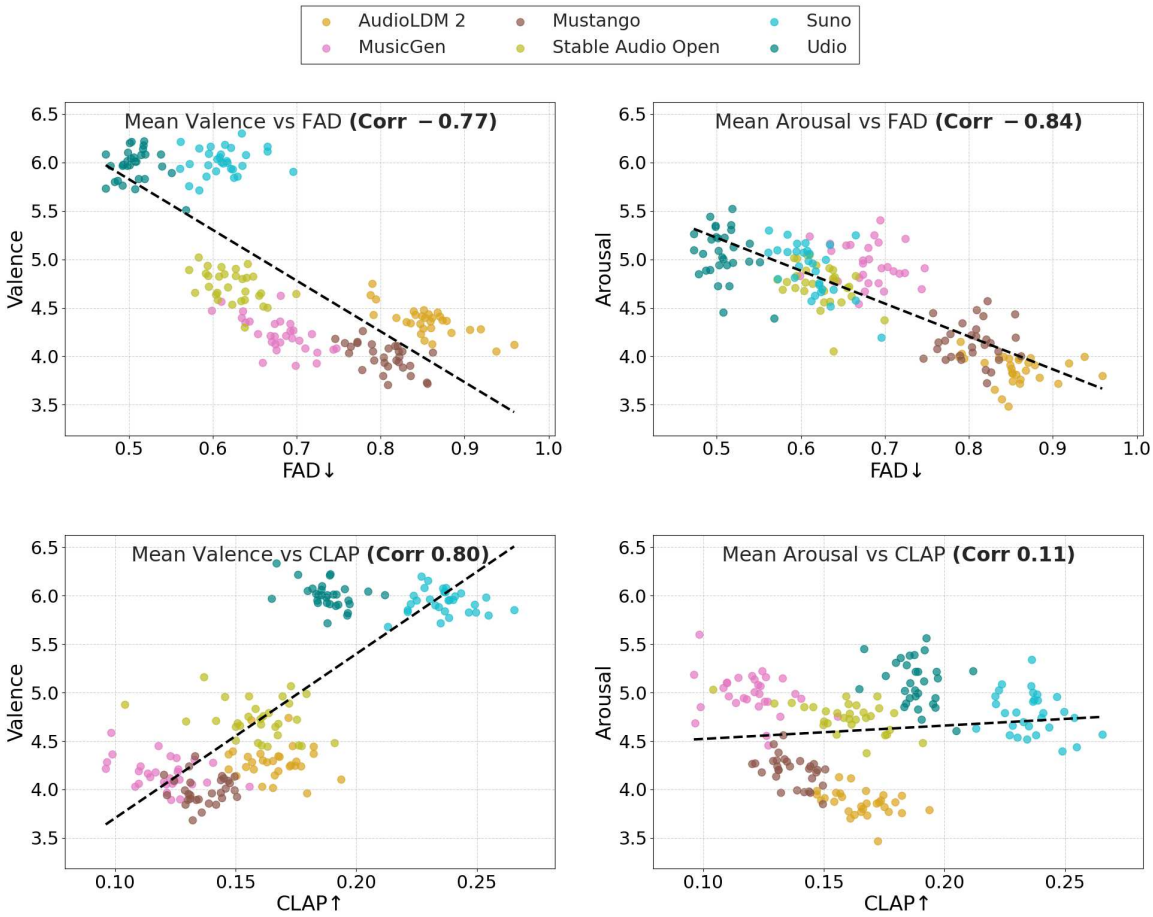


Figure 4. Scatter plots illustrating correlations between bootstrapped valence-arousal clip ratings and automatic metrics. (upper: FAD, lower: CLAP score) Pearson correlation coefficients are calculated for each combination of affective dimensions and automatic metrics.

Correlation between Automatic Metrics and Affective Ratings

To examine the correlation between clip generation-related factors and variations in emotional conveyance, we evaluated the generated audio quality and text alignment utilizing two objective metrics: Fréchet Audio Distance (FAD) and CLAP score.

- **FAD (Fréchet Audio Distance)**[32]: This metric assesses overall music quality by measuring the distribution difference between the embeddings of a reference dataset and the generated audio. Lower FAD scores typically correlate with higher audio quality. We utilized LAION-CLAP-MA[36] (`checkpoint:music_audioset_epoch_15_esc_90_14.pt`) as the embedding model. The reference distribution was created using 2,000 randomly selected tracks from the MTG-Jamendo dataset that consists of 56,639 high-quality music samples[37] due to resource constraints. It was shown that scores from LAION-CLAP-MA with the MTG-Jamendo dataset as a reference dataset align closely with human preferences[15].
- **CLAP Score**[33]: This metric measures the semantic alignment between the text prompt (emotion intent) and the generated audio utilizing CLAP, which projects both text and audio into a unified latent space. It is calculated as the cosine similarity between the text and audio embeddings within the shared CLAP latent space. Higher scores indicate better alignment with the intended prompt. We employed LAION-CLAP-MA, the variant of CLAP that was shown to exhibit the strongest correlation with human preference[15].

FAD scores exhibited strong negative correlations with both perceived valence and arousal. The correlation was particularly robust for arousal ($r = -0.84$, 95% CI $[-0.88, -0.79]$) compared to valence ($r = -0.77$, 95% CI $[-0.81, -0.73]$). These Pearson correlations are descriptive statistics computed on bootstrapped means (sampling 50 clips over 30 iterations per system); confidence intervals were obtained by resampling the bootstrapped data points with replacement, because FAD is a

distribution-based metric that requires multiple samples to calculate. A Linear Mixed Model (LMM) with FAD as fixed effect and emotion intent as random intercept was used for statistical test, and showed statistical significance of the FAD effect on both valence ($\beta = -0.574, p < .001$) and arousal ($\beta = -0.433, p < .001$) ratings. Residual diagnostics confirmed linearity of the predictor-residual relationship; non-normality (Shapiro-Wilk: valence $W = 0.988, p < .05$, arousal $W = 0.995, p < .05$) and heteroscedasticity (Breusch-Pagan: valence $LM = 18.70, p < .05$, arousal $LM = 31.90, p < .05$) were observed in the residuals, though the large sample size of ratings ($n = 6{,}162$) ensures the robustness of the LMM inference.

Similarly, CLAP scores showed a strong positive correlation with valence ($r = +0.80$, 95% CI $[0.75, 0.84]$) but a negligible correlation with arousal ($r = +0.11$, 95% CI $[0.00, 0.23]$). LMM analysis with CLAP score as fixed effect and emotion intent as a random intercept further revealed that while the CLAP score was a significant predictor for valence ratings ($\beta = 5.743, p < .001$), it had no significant impact on perceived arousal ($\beta = -0.116, p = .746$). Residual diagnostics for the CLAP LMMs similarly confirmed linearity, with non-normality (valence: $W = 0.987, p < .05$, arousal: $W = 0.993, p < .05$) and heteroscedasticity (Breusch-Pagan: valence $LM = 130.49, p < .05$, arousal $LM = 78.52, p < .05$) that can be mitigated by the large sample size. These results indicate that while overall audio quality affects both valence and arousal, semantic alignment to the text mainly relates to the communicated valence level in TTM systems.

Discussion

This study provides a systematic and perceptually grounded evaluation of emotion conveyance in text-to-music (TTM) systems. By adopting the valence–arousal framework[23,24] and collecting two-dimensional ratings from human listeners, we move beyond existing methods using coarse categorical evaluations[31]. We numerically characterize how generated music aligns with intended emotion, using normative valence–arousal ratings of the emotion words employed in the prompts[38]. This design enables fine-grained analyses of quadrant-specific performance, rating distributions, and system-dependent biases that are difficult to capture with purely categorical schemes.

Our approach focuses on systems' tendency to convey emotional intensity and direction and how this tendency differs across models, highlighting strengths and weaknesses in the valence-arousal space. Furthermore, by utilizing objective evaluation metrics (FAD and CLAP) and affective theories such as BRECVEMAC[39], we offer a grounded interpretation of the results. Moreover, we release AImoclips, an open-source dataset generated by multiple TTM systems with corresponding human-rated valence–arousal scores. This dataset supports reproducible benchmarking and further investigations into TTM emotion conveyance, perceptual variability, and the relationship between generative modeling choices and human affective responses.

The most notable finding of this study is that current TTM systems exhibit model-dependent differences in emotion conveyance across the VA dimensions. In particular, commercial systems exhibited substantially higher quadrant-level performance in high-valence quadrants (Table 2). TTM systems exhibited biases in distinct directions, allowing them to be grouped into three affective clusters: Suno and Udio have higher valence and higher arousal; AudioLDM 2 and Mustango have lower valence and lower arousal; and MusicGen and Stable Audio Open have lower valence and higher arousal. Such emotion-specific generation biases have also been reported in prior work on emotion-aware music generation and recognition[40], suggesting that both training data and post-processing strategies may cause systematic biases toward certain affective regions.

To further examine potential factors underlying these differences, we analyzed correlations between VA ratings and FAD, which could explain the common tendency of commercial systems. FAD had strong negative correlations with both valence and arousal, indicating that a higher preference score (lower FAD) was associated with higher affective intensity along both dimensions. Although the original FAD has been reported to correlate only moderately with human judgments of music quality[32], the specific FAD setting used in this study was shown to achieve improved alignment with human preference[15]. In addition, music produced by commercial systems (Suno and Udio) showed lower FAD values than that produced by open-source systems (Figure 4), which aligns with earlier findings that Suno and Udio achieve higher human preference ratings than several open-source systems (AudioLDM 2, Mustango, and MusicGen) used in this study[15]. Taken together, these findings provide a plausible explanation for the positive valence and arousal deviations observed in commercial models (Figure 2), suggesting that listeners may attribute stronger pleasurable affect to highly preferred outputs, consistent with prior work[41].

In addition, the analysis of CLAP scores further strengthens the positive valence bias. There was a positive correlation between CLAP scores and valence, but little correlation with arousal. Since CLAP measures text–audio alignment, the correlation suggests that systems with stronger semantic alignment tend to produce outputs perceived as more positively valenced. As it has been shown that Suno and Udio outperform most other open-source systems in text-audio alignment[15], their ability to match the music output with the intended text can also help explain why commercial systems show a positive valence deviation. However, text-audio alignment cannot account for positive arousal deviations due to the lack of correlation between CLAP scores and arousal ratings.

'Pleasurable Sadness'[42] can be a complementary explanation for the high-valence bias observed in Suno and Udio, which did not appear in open-source systems. Prior research has shown that sadness expressed in music can elicit pleasurable or

aesthetically rewarding experiences, distinct from everyday sadness[43–45], and higher musical and aesthetic quality can further amplify enjoyment[46]. Within Juslin's BRECVEMAC framework, aesthetic judgment (component A) can override core affective responses, leading listeners to experience negatively valenced emotions as more pleasant[47]. Accordingly, relatively high-quality music generated by commercial systems may cause negative feelings to seem more pleasant, leading to more accurate ratings in Q1 and Q4, but not as much in Q2 and Q3.

Conversely, the low-arousal bias observed in AudioLDM 2 and Mustango may stem from a lack of musically salient structure. Unlike MusicGen and Stable Audio Open, many generated samples exhibited limited rhythmic or dynamic variation, often characterized by sustained tonal textures that obscured clear melodic or structural progression, which was also reflected in clearly higher FAD scores for those systems than for other systems (Figure 4). In Juslin's model, the brain stem reflex (component B) is triggered by salient acoustic events and dynamic changes, which cause elevated arousal[47]. Moreover, it was shown that musical features such as event density and timbral clarity are strongly predictive of perceived energy and tension[48], indicating that structural salience plays a central role in affective communication. Therefore, clips that lack musically notable events would reduce perceived arousal, resulting in system-level arousal bias.

MusicGen and Stable Audio Open show positive arousal bias, like Suno and Udio, while the direction of valence bias is the same as in other open-source systems. Although MusicGen and Stable Audio Open show comparable FAD to commercial systems, their conveyed valence was lower than intended, unlike commercial ones. This difference could result from the limited range of musical features that FAD can reflect. While FAD highly correlates with audio fidelity or human preference (in the setting we used), it does not comprehensively encapsulate musical elements such as mode and key, which are believed to be closely linked to valence[42]. Because the reference dataset contains music with diverse tonal properties, FAD may not sensitively reflect specific variations in mode or key. Consequently, the resultant negative valence bias of MusicGen and Stable Audio Open may indicate a tendency to produce songs with tonal characteristics perceived as less pleasant.

Beyond model-specific differences, our results reveal several global tendencies shared across TTM systems. At the quadrant level, all systems conveyed quadrant-wise emotional intent at rates significantly above chance. This result highlights an important distinction: while current TTM systems do play a role with respect to emotion conveyance, their performance remains insufficient for reliable emotion control in practical applications. Two additional findings are noteworthy: Q1 (high valence–high arousal) consistently achieved the highest performance across all metrics, and a large proportion of Q4 targets were rated as Q3. Global clip rating distribution also demonstrated a noticeable feature: the centralizing tendency.

The overall performance measurement matches earlier results[31], which looked at "emotion conversion accuracy" using a different experimental setup. Unlike the present study, which transformed dimensional VA ratings into quadrant-level categories, their evaluation employed a multiple-choice task in which participants selected one of four emotion words representing each VA quadrant. Under this evaluation scheme, they reported accuracies of 0.604 for MusicLM, 0.530 for Stable Audio, and 0.436 for MusicGen. Although these values are higher than the accuracy calculated here, they remain far from ceiling performance and are not directly comparable due to differences in task formulation and error criteria. The present study's lower accuracy likely stems from a stricter evaluation that penalizes small perceptual shifts across quadrant boundaries as errors.

The positive correlation between CLAP score and rated valence provides a potential explanation for the overall tendency of TTM systems to perform better in Q1. As mentioned before, a higher CLAP score indicates higher text-audio alignment, which reflects stronger semantic alignment between the textual prompt and the generated audio. These results can be related to more accurate emotion conveyance, as the prompt was only constructed with a single word representing emotion intent, except for the instrument-only control prompt. Therefore, the positive correlation indicating an intimate relationship between CLAP score and perceived valence would indicate that clips with higher valence ratings tend to be associated with high semantic alignment of emotion.

Figure 1(A) further shows substantial overlap between Q3 and Q4 score distributions across most models, suggesting that low-arousal affective boundaries are intrinsically difficult to determine for AI-generated music. This explanation is consistent with prior music emotion recognition (MER) studies demonstrating that the emotions 'calm' and 'sad' are difficult to distinguish in music[49] and that listener recognition accuracy is generally higher for high-arousal emotions than for low-arousal ones[50]. From a cue-based perspective, while happy and sad emotions are often distinguished by the canonical contrast of musical cues (e.g., major/fast versus minor/slow)[51,52], peaceful/calm and sad emotions share many acoustic features, such as slow tempo, soft dynamics, and smooth articulation, which leads to boundary ambiguity. This ambiguity helps explain why performance in Q4 remained low, unlike Q1, despite CLAP-based semantic alignment suggesting that both quadrants had the potential for accurate emotion conveyance.

A discrepancy arises from the comparison between categorical results and the distances between rated scores and reference scores. Despite Q1 showing the highest accuracy, its mean distance from reference scores is greater than that of the other quadrants, while Q2 exhibits the smallest distance. This pattern can be understood by jointly considering the distributions in the VA plane (Figure 1(A)) and the reference score locations. For Q1, the reference scores lie at an extremely high valence level, while many clips tend to cluster near the center, showing a moderate valence and arousal tendency. This results in many Q1

clips remaining within the correct quadrant with high categorical accuracy, yet located far from reference points. In contrast, Q2 reference scores show less extreme valence. Although clips from Q2 display wide variability across quadrants, resulting in low categorical performance, many of them tend to fall closer to the reference location. Taken together, these results indicate that categorical accuracy and distance distribution capture complementary aspects of emotion conveyance.

Finally, across all systems, rated clips tended to cluster closer to the neutral center than the corresponding reference scores. This centralizing tendency was statistically significant and suggests that valence and arousal expressions in generated music are generally less distinct than intended. A key factor lies in the fundamental difference between semantic communication in language and in music. Emotion words used as prompts have relatively stable and well-defined semantic meanings, while music communicates in a more probabilistic, affective, and ambiguous manner[53,54]. As a result, when people listen to music and interpret musical emotion, a single clip is plausible to convey multiple affective meanings, varying among listeners in contrast to textual words. This communicative ambiguity leads to greater variability in individual ratings, resulting in aggregated clip-wise scores that exhibit attenuated valence–arousal intensity.

A limitation of this study was the limited number of ratings per clip. The AImoclips dataset includes ratings from 111 participants across 991 generated clips, resulting in approximately 4–9 ratings per clip. While this scale is sufficient to reveal robust aggregate trends, the relatively small number of raters per clip may reduce statistical power for fine-grained analyses and contribute to lower inter-rater agreement. Future studies should benefit from substantially increasing both the total number of clips and the number of ratings per clip to enable more robust analysis of emotion conveyance.

Another limitation concerns the reference valence–arousal scores used for emotion intents, which introduce potential sources of mismatch. The reference scores and the clip ratings were obtained from different groups of participants. Although the same numerical scale was employed, its internal calibration may differ across groups. Because the meaning and expression of emotions in music can differ from those in language, some discrepancy may arise between the intended textual emotion and the perceived emotion in music. Future work would consider a more controlled design, such as collecting reference ratings for emotion words and clip ratings from the same participants, which could reduce this source of variability.

Extending the participant group and survey design, and broadening the analyses could further strengthen this line of research. Because all participants in this study were Korean, the results may reflect cultural biases; recruiting raters from more diverse cultural backgrounds would improve generalizability. Moreover, participants rated only valence and arousal on a numerical scale, without providing categorical emotion labels or explicit judgments of music quality or preference. Consequently, categorical interpretations had to be inferred via quadrant-based mappings of VA scores, and links between perceived affect and preference were examined indirectly using objective metrics such as FAD. Future survey designs could add multiple-choice emotion labels and direct quality and preference ratings. Collecting these measures from the same raters who provide VA scores would enable more direct and comprehensive comparisons among perceived emotion, semantic intent, subjective preference, and audio quality.

The rapid evolution of TTM systems has led to the continual release of new foundation models[55,56]. It remains an open question whether the quadrant-specific biases and perceptual tendencies observed here will persist in newer systems with improved coherence and music quality. Future research should therefore update and extend this benchmark to include emerging models and architectures, allowing long-term analyses of progress in emotion conveyance. In addition, emotions induced or perceived in music are known to be domain-specific and may not be fully representative within a low-dimensional affective space[53,54,57]. Expanding evaluation beyond the valence–arousal framework to richer affective models may therefore capture aspects of emotion conveyance that VA-based metrics alone cannot reflect.

To conclude, this study provides a perceptually grounded benchmark of emotion conveyance in text-to-music systems. While listeners could meaningfully perceive coarse emotional intent from AI-generated music, its emotion-conveying ability was insufficient for stable emotional control. Moreover, it was uneven across the valence-arousal space and varied across different systems. These results highlight a fundamental challenge in emotion-aware music generation: translating semantically precise linguistic emotion intent into musically conveyed affect that is both controllable and perceptually consistent. Addressing this challenge will require advances in both generative modeling and affective representations that better reflect how humans experience emotion in music.

Methods

Emotion Intents

To express emotional intent in music clips, we initially selected 12 emotion words. Prior to the final selection, we utilized two sets of words as potential candidates. The initial set comprises 28 sample words for the Circumplex of Affect model[23]. The second set comprises adjectives from the 12-Point Affect Circumplex Model (12-PAC)[58]. To exclude words with neutral valence or arousal, we only incorporated terms not lying on the valence or arousal axes in this collection. We selected terms from these sets with valence and arousal scores documented in Warriner et al.[34], an extension of Affective Norms for English Words (ANEW)[35], and excluded words with valence or arousal scores ranging from 4.0 to 6.0 on a scale of 1.0 to 9.0. This step

was designed to exclude words with intermediate valence or arousal values to clearly differentiate the groups of emotional intents associated with each quadrant of the valence-arousal plane. Finally, we utilized 5.0 (for both valence and arousal) as the threshold and selected three words for each quadrant. Table 3 presents a compilation of the selected words.

Table 3. Words chosen for emotion intent of TTM generation

Quadrant	Valence	Arousal	Words
Q1	High	High	happy, excited, energetic
Q2	Low	High	angry, anxious, scared
Q3	Low	Low	sad, gloomy, dull
Q4	High	Low	relaxed, calm, tranquil

TTM System Selection and Clip Generation

Subsequently, we selected six TTM systems and generated 14 clips for each system and emotional intent. A dataset comprising 1,008 AI-generated music clips is created utilizing 12 emotional intents, resulting in 168 clips per system. Each clip had a duration of 10 seconds. We employed two commercial models, Suno v4.5[13] and Udio v1.5 Allegro[14], alongside four open-source models: AudioLDM 2[59], MusicGen[8], Mustango[9], and Stable Audio Open[10]. Diffusion-based[60] generative audio models encompass AudioLDM 2 and Stable Audio Open. MusicGen is an autoregressive transformer-based[61] decoder model, while Mustango is a hybrid model that integrates a diffusion model with an instruction-tuned language model, FLAN-T5[62]. Table 4 presents a summary of each system accompanied by the sample rate. We could not standardize the sample rate, as it was predetermined for each checkpoint.

Table 4. Information of TTM systems to generate music clips in AImoclips

Model	Checkpoint (Version)	Availability	Sample rate(kHz)
AudioLDM 2	audioldm2-music	open-source	16
MusicGen	musicgen-medium	open-source	32
Mustango	mustango	open-source	16
Stable Audio Open	stable-audio-open-1.0	open-source	44.1
Suno	Suno v4.5	commercial	44.1
Udio	Udio v1.5 Allegro	commercial	44.1

All generations adhered to a uniform prompt format. The text prompt format “word, instrumental” was used with each TTM system to generate music corresponding to a specific emotional intent word. The inclusion of "instrumental" in the prompt was intended to guarantee the generation of music clips without vocals. Given that vocal lyrics can affect the emotional expression of the clip, we aimed to examine the impact of the music alone, excluding any vocals. Suno and Udio feature an "instrumental" toggle option to inhibit the models from producing vocal tracks. Consequently, the text input solely comprises the emotion intent (“word”) without supplementary words, and the “instrumental” option is activated.

Following the creation of original audio files with each TTM system, a random cropping method was employed to standardize the clip duration to 10 seconds. This is because the initial clips produced by Suno and Udio exceeded a duration of 1 minute. Each clip was initially generated with a duration of 30 seconds for open-source models, followed by a random extraction of a 10-second segment from the 15% to 85% range of the complete clip to omit the introduction and conclusion, which generally contain minimal information. It was unfeasible to precisely control the output length produced by Suno and Udio; therefore, we implemented an identical cropping procedure, with the only difference being that the full-length clips exhibited variable lengths. All clips had a minimum duration of 30 seconds, adequate for extracting a 10-second segment.

Human Evaluation

An online survey was conducted to gather valence-arousal ratings for the produced music clips. Each of the 111 participants rated 56 randomly selected clips, yielding 6,162 ratings that surpassed five times the total number of clips (1,008), averaging over five ratings per clip. Each clip was chosen from a unique combination of TTM systems utilized for generation and emotional intent, subsequently presented in a random sequence. Participants assessed the expressed valence and arousal using a 9-point Likert scale. All evaluation activities involving human subjects in this study received approval from the Institutional Review Board at Seoul National University, under approval number IRB No. 2505/001-012, and informed consent was obtained

from all participants. Additionally, we confirm that all methods in this study were carried out in accordance with relevant guidelines and regulations.

Prior to evaluating the clips, participants were requested to provide demographic details, including gender and age. This procedure ensured that all participants were proficient Korean speakers aged 18 or older. Subsequently, they were presented with a sample question (identical to those in the primary survey, excluding the evaluation of valence and arousal) and a clip to verify their auditory comprehension and understanding of the instructions. The sample clip was excluded from the dataset. The sole additional data gathered was contact information intended for the distribution of a participation reward.

Human evaluation was carried out using the Modified Self-Assessment Manikin (SAM) scale. Participants were shown each clip sequentially and were asked two questions (in Korean): "Rate the pleasure conveyed by this music clip, with 1 indicating 'very unpleasant' and 9 indicating 'very pleasant,'" and "Rate the arousal conveyed by this music clip, with 1 indicating 'very relaxed' and 9 indicating 'very aroused.'" To assist participants in understanding the concepts of valence (pleasure) and arousal, the images in Figure 1 were presented alongside both questions. This approach integrates visual aids with the explicit textual anchors present in the questions. Participants assigned integer scores ranging from 1 to 9 for each question and were required to submit a single rating per question to advance to the subsequent clip, akin to the static annotation method employed in the DEAM dataset[19]. All valence and arousal ratings from participants who responded to all 56 clips were incorporated into the dataset and averaged per clip.

In supplying music clips for a new participant, clips evaluated by a limited number of participants were given higher priority to balance the rating distribution for each clip. Nonetheless, achieving a complete balance in the ratings was shown to be challenging, as certain participants failed to finish the survey. The number of participants rating each clip ranged from 2 to 9, leading to the exclusion of 17 clips with three or fewer ratings, resulting in a final dataset of 991 clips. The quantity of ratings for each clip varies from 4 to 9.

Data Availability

The datasets generated during the current study are available in the 'HunRotation/AImoclips' GitHub repository, https://github.com/HunRotation/AImoclips.

References

1. van den Oord, A. *et al.* WaveNet: A Generative Model for Raw Audio. *arXiv* DOI: 10.48550/arXiv.1609.03499 (2016). 1609.03499.
2. Huang, C.-Z. A. *et al.* Music Transformer: Generating Music with Long-Term Structure. In *International Conference on Learning Representations* (2018).
3. Huang, Y.-S. & Yang, Y.-H. 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, DOI: 10.1145/3394171.3413671 (2020).
4. Roberts, A., Engel, J., Raffel, C., Hawthorne, C. & Eck, D. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the 35th International Conference on Machine Learning*, 4364–4373 (2018).
5. Payne, C. Musenet. https://openai.com/blog/musenet (2019). OpenAI blog post.
6. Dhariwal, P. *et al.* Jukebox: A generative model for music. *arXiv* DOI: 10.48550/arXiv.2005.00341 (2020). 2005.00341.
7. Agostinelli, A. *et al.* MusicLM: Generating Music From Text. *arXiv* DOI: 10.48550/arXiv.2301.11325 (2023). 2301.11325.
8. Copet, J. *et al.* Simple and Controllable Music Generation. *Adv. Neural Inf. Process. Syst.* **36**, 47704–47720, DOI: 10.5555/3666122.3668188 (2023).
9. Melechovsky, J. *et al.* Mustango: Toward Controllable Text-to-Music Generation. In Duh, K., Gomez, H. & Bethard, S. (eds.) *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8293–8316, DOI: 10.18653/v1/2024.naacl-long.459 (2024).
10. Evans, Z. *et al.* Stable Audio Open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, DOI: 10.1109/ICASSP49660.2025.10888461 (2025).
11. Liu, S., Hussain, A. S., Wu, Q., Sun, C. & Shan, Y. M^{2}UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv* DOI: 10.48550/arXiv.2311.11255 (2024). 2311.11255.
12. Schneider, F., Kamal, O., Jin, Z. & Schölkopf, B. Moûsai: Efficient Text-to-Music Diffusion Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8050–8068, DOI: 10.18653/v1/2024.acl-long.437 (2024).

13. Suno. Suno. https://suno.com/ (2024). Accessed: May-June, 2025.

14. Udio. Udio. https://www.udio.com/ (2024). Accessed: May-June, 2025.

15. Grötschla, F., Solak, A., Lanzendörfer, L. A. & Wattenhofer, R. Benchmarking Music Generation Models and Metrics via Human Preference Studies. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, DOI: 10.1109/ICASSP49660.2025.10887745 (2025).

16. Liu, C. *et al.* MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, DOI: 10.1109/ICASSP49660.2025.10890307 (2025).

17. Huang, Y. *et al.* Aligning Text-to-Music Evaluation with Human Preferences. *arXiv* DOI: 10.48550/arXiv.2503.16669 (2025). 2503.16669.

18. Hevner, K. Experimental Studies of the Elements of Expression in Music. *The Am. J. Psychol.* **48**, 246–268, DOI: 10.2307/1415746 (1936). 1415746.

19. Aljanaki, A., Yang, Y.-H. & Soleymani, M. Developing a benchmark for emotional analysis of music. *PLOS ONE* **12**, e0173392, DOI: 10.1371/journal.pone.0173392 (2017).

20. Turnbull, D., Barrington, L., Torres, D. & Lanckriet, G. Towards musical query-by-semantic-description using the CAL500 data set. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, 439–446, DOI: 10.1145/1277741.1277817 (2007).

21. Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y. & Yang, Y.-H. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, 1–6, DOI: 10.1145/2506364.2506365 (2013).

22. Eerola, T. & Vuoskoski, J. K. A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music.* **39**, 18–49, DOI: 10.1177/0305735610362821 (2011).

23. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178, DOI: 10.1037/h0077714 (1980).

24. Posner, J., Russell, J. A. & Peterson, B. S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **17**, DOI: 10.1017/S0954579405050340 (2005).

25. Hung, H.-T. *et al.* EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. *arXiv* DOI: 10.48550/arXiv.2108.01374 (2021). 2108.01374.

26. Ferreira, L. N. & Whitehead, J. Learning to Generate Music With Sentiment. *arXiv* DOI: 10.48550/arXiv.2103.06125 (2021). 2103.06125.

27. Grekow, J. & Dimitrova-Grekow, T. Monophonic Music Generation With a Given Emotion Using Conditional Variational Autoencoder. *IEEE Access* **9**, 129088–129101, DOI: 10.1109/ACCESS.2021.3113829 (2021).

28. Choi, E. *et al.* YM2413-MDB: A Multi-Instrumental FM Video Game Music Dataset with Emotion Annotations. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 100–108 (2022).

29. Sulun, S., Davies, M. E. P. & Viana, P. Symbolic Music Generation Conditioned on Continuous-Valued Emotions. *IEEE Access* **10**, 44617–44626, DOI: 10.1109/ACCESS.2022.3169744 (2022).

30. Huang, J., Chen, K. & Yang, Y.-H. Emotion-driven Piano Music Generation via Two-stage Disentanglement and Functional Representation. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR 2024)*, 335–342 (2024).

31. Gao, X. *et al.* AI-driven music generation and emotion conversion. *AHFE Int.* DOI: 10.54941/ahfe1004679 (2024).

32. Kilgour, K., Zuluaga, M., Roblek, D. & Sharifi, M. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*, 2350–2354, DOI: 10.21437/Interspeech.2019-2219 (2019).

33. Elizalde, B., Deshmukh, S., Ismail, M. A. & Wang, H. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, DOI: 10.1109/ICASSP49357.2023.10095889 (2023).

34. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207, DOI: 10.3758/s13428-012-0314-x (2013).

35. Bradley, M. M. & Lang, P. J. Instruction Manual and Affective Ratings. Tech. Rep. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida (1999).

36. Wu, Y. *et al.* 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, DOI: 10.1109/ICASSP49357.2023.10095969 (2023).

37. Bogdanov, D., Won, M., Tovstogan, P., Porter, A. & Serra, X. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop (ML4MD) at the International Conference on Machine Learning (ICML 2019)* (Long Beach, CA, USA, 2019).

38. Mehrabian, A. & Russell, J. A. *An Approach to Environmental Psychology*. An Approach to Environmental Psychology (The MIT Press, 1974).

39. Juslin, P. N. Major theories of emotion causation and their applicability to music: The case for multi-level approaches. *Music. Perception: An Interdiscip. J.* **42**, 421–466 (2025).

40. Li, Y., Gui, A., Emmanouilidou, D. & Gamper, H. Addressing Emotion Bias in Music Emotion Recognition and Generation with Frechet Audio Distance. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6, DOI: 10.1109/ICME59968.2025.11210239 (2025).

41. Fuentes-Sánchez, N., Pastor, R., Eerola, T., Escrig, M. A. & Pastor, M. C. Musical preference but not familiarity influences subjective ratings and psychophysiological correlates of music-induced emotions. *Pers. Individ. Differ.* **198**, 111828, DOI: https://doi.org/10.1016/j.paid.2022.111828 (2022).

42. Juslin, P. N. & Sloboda, J. A. *Handbook of Music and Emotion: Theory, Research, Applications* (OUP Oxford, 2010).

43. Kawakami, A., Furukawa, K., Katahira, K. & Okanoya, K. Sad music induces pleasant emotion. *Front. Psychol.* **4**, DOI: 10.3389/fpsyg.2013.00311 (2013).

44. Sachs, M. E., Damasio, A. & Habibi, A. The pleasures of sad music: A systematic review. *Front. Hum. Neurosci.* **9**, DOI: 10.3389/fnhum.2015.00404 (2015).

45. Eerola, T., Vuoskoski, J. K., Peltola, H.-R., Putkinen, V. & Schäfer, K. An integrative review of the enjoyment of sadness associated with music. *Phys. Life Rev.* **25**, 100–121, DOI: 10.1016/j.plrev.2017.11.016 (2018).

46. Eerola, T., Vuoskoski, J. K. & Kautiainen, H. Being Moved by Unfamiliar Sad Music Is Associated with High Empathy. *Front. Psychol.* **7**, DOI: 10.3389/fpsyg.2016.01176 (2016).

47. Juslin, P. N. From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Phys. Life Rev.* **10**, 235–266, DOI: 10.1016/j.plrev.2013.05.008 (2013).

48. Pring, E. X., Olsen, K. N., Mobbs, A. E. D. & Thompson, W. F. Music communicates social emotions: Evidence from 750 music excerpts. *Sci. Reports* **14**, 27766, DOI: 10.1038/s41598-024-78156-1 (2024).

49. Hong, Y., Chau, C.-J. & Horner, A. An Analysis of Low-Arousal Piano Music Ratings to Uncover What Makes Calm and Sad Music So Difficult to Distinguish in Music Emotion Recognition. *J. Audio Eng. Soc.* **65**, 304–320, DOI: 10.17743/jaes.2017.0001 (2017).

50. Paz, Y. *et al.* Hearing a Feeling: Music Emotion Recognition and Callous-Unemotional Traits in Early Childhood. *Child Dev.* **96**, 2247–2256, DOI: 10.1111/cdev.70024 (2025).

51. Peretz, I., Gagnon, L. & Bouchard, B. Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition* **68**, 111–141, DOI: 10.1016/S0010-0277(98)00043-2 (1998).

52. Brattico, E. *et al.* A Functional MRI Study of Happy and Sad Emotions in Music with and without Lyrics. *Front. Psychol.* **2**, DOI: 10.3389/fpsyg.2011.00308 (2011).

53. Meyer, L. B. *Emotion and Meaning in Music* (University of Chicago Press, 1956).

54. Cespedes-Guevara, J. & Eerola, T. Music Communicates Affects, Not Basic Emotions – A Constructionist Account of Attribution of Emotional Meanings to Music. *Front. Psychol.* **9**, DOI: 10.3389/fpsyg.2018.00215 (2018).

55. Gong, J., Zhao, S., Wang, S., Xu, S. & Guo, J. ACE-Step: A Step Towards Music Generation Foundation Model. *arXiv* DOI: 10.48550/arXiv.2506.00045 (2025). 2506.00045.

56. Yang, D. *et al.* HeartMuLa: A Family of Open Sourced Music Foundation Models. *arXiv* DOI: 10.48550/arXiv.2601.10547 (2026). 2601.10547.

57. Zentner, M., Grandjean, D. & Scherer, K. R. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion* **8**, 494–521, DOI: 10.1037/1528-3542.8.4.494 (2008).

58. Yik, M., Russell, J. A. & Steiger, J. H. A 12-point circumplex structure of core affect. *Emotion* **11**, 705–731, DOI: 10.1037/a0023980 (2011).

59. Liu, H. *et al.* AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, Lang. Process.* **32**, 2871–2883, DOI: 10.1109/TASLP.2024.3399607 (2024).

60. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, vol. 33, 6840–6851 (2020).

61. Vaswani, A. *et al.* Attention is All you Need. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

62. Chung, H. W. *et al.* Scaling Instruction-Finetuned Language Models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).

63. Go, G. *et al.* AImoclips: A benchmark for evaluating emotion conveyance in text-to-music generation. *arXiv* DOI: 10.48550/arXiv.2509.00813 (2025). 2509.00813.

Acknowledgements

Preliminary analysis of AImoclips dataset was presented at a workshop[63] as an oral presentation.

Author contributions statement

G.G., S.H., A.C., E.C., J.N., and J.M.P. contributed to the conception and design of research. G.G. and S.H. collected the data. G.G., S.H., and J.M.P. analyzed and interpreted the results. G.G. wrote the first draft of the manuscript. S.H., A.C., E.C., J.N., and J.M.P. provided suggestions and revisions on the manuscript.

Funding

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5B5A17087935).

Ethics declarations

This study received approval from the Institutional Review Board at Seoul National University, under approval number IRB No. 2505/001-012.

Approval for human experiments

Informed consent was obtained from all participants.

Additional information

Competing interests The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AlmoclipsSciRepSupplementary.pdf