# LLM-Guided Affective Music Generation via Prompt-to-Parameter Mapping and Frequency-Aware Evaluation

Bingxu Jin[1, a]

[1]Beijing International Bilingual Academy, Beijing, China

[a]2027kjin@biba-student.org

## Abstract

**Recent text-to-music generation models have enabled users to compose music from natural language description; they typically regard emotional prompts as a black-box conditioning signal, which is hard to interpret, control, and assess. In this paper, we present an LLM-guided affective music creation framework featuring prompt-to-parameter transformation and frequency-aware evaluation. Our approach begins by taking a handwritten emotional cue and applying a large language model as affective mapper that converts subjective language into structured music parameters such as tempo, dynamics, harmony, texture density, register, instrumentation and spectral targets. These parameters are then input into a downstream model that generates music. For music generation evaluation, we propose a multi-dimensional protocol on the aspects of audio fidelity, text-music alignment, and frequency-domain consistency. We evaluate 40 affective prompts, and our method achieves 0.536 of CLAP alignment, 4.31 of FAD, and 0.792 of spectral consistency, while direct prompt generation achieves 0.412, 5.87, and 0.641, respectively. Subjective listening results also reveal that our method results in higher overall quality, emotional consistency and listening comfort. These results show that the explicit LLM-based affective mapping offers a more interpretable and controllable approach to emotional music generation.**

## Keywords

**Text-To-Music Generation; Affective Music Generation; Large Language Models; Prompt-To-Parameter Mapping; Frequency-Aware Evaluation.**

## 1. Introduction

Recent advances in generative modeling have substantially changed the way music can be composed, performed, and evaluated. Large-scale text-conditioned audio and music generation models, such as MusicLM [1], MusicGen [3], and AudioLDM [8], have demonstrated impressive capability in synthesizing coherent musical audio from natural language descriptions. In parallel, audio-language representation learning methods, including MuLan [6] and CLAP [5], have enabled cross-modal comparison between textual descriptions and audio signals. These developments suggest a new paradigm in which natural language is no longer only used for metadata retrieval, but becomes an expressive interface for music creation.

While there has been significant improvement, existing text-to-music systems have not yet overcome several challenges in the context of affective and controllable music generation. First, free-form text prompts often include rich emotional, contextual, and metaphorical descriptions, such as "tired but hopeful, like walking home in light rain", which the existing generation models often take as opaque conditions. The intermediate intentions of music in respect of tempo, dynamics, harmonic tension, register, texture density, and/or spectral brightness are not often stated. This makes the generation process hard to interpret, modify or satisfy a composer's wishes. Second, while most systems have concentrated on perceptual quality or on

similarity between text and audio, they have had little analysis of the properties of the frequency domain that are essential for musical emotion and timbre, or for application in real acoustic environments. Third, there is often a problem with the fragmentation of the evaluation process: audio fidelity, semantic alignment, and spectral behavior are treated separately, and subjective affective intention and objective acoustic evidence do not align.

In this paper, an affect-aware music creation framework is proposed based on frequency analysis and the use of a large language model for affect mapping. Our method relies on a large language model (LLM) as an interpretable emotional translator instead of directly inputting a raw prompt into a music generation model. The LLM translates the affective description into a set of musical control parameters, such as rhythm, dynamics, harmonic color, texture density, register, instrumentation tendency, and frequency-related targets, given a handwritten natural language prompt. These parameters are then fed to an audio model at the output end to generate music that is not only text-guided, but also explicitly controlled by musically meaningful descriptions.

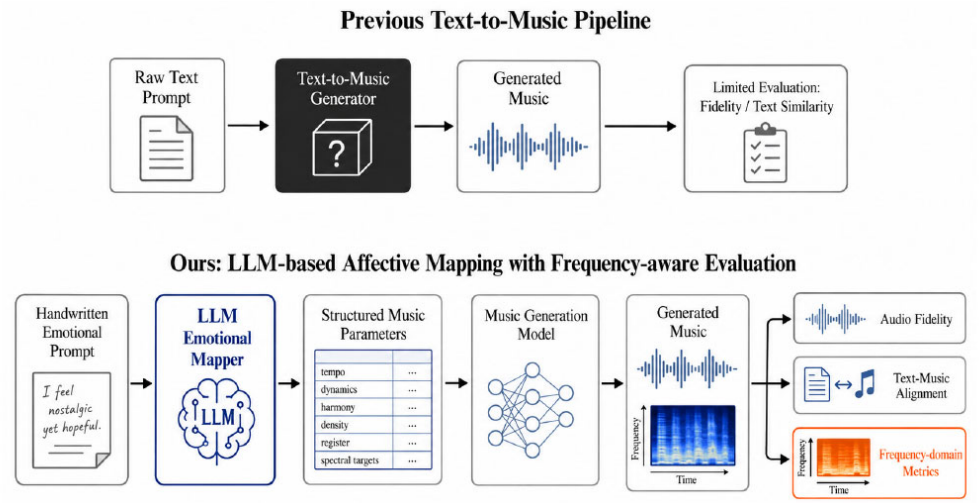


**Figure 1.** Comparison between the previous text-to-music pipeline and our proposed framework.

We also present a multi-dimensional evaluation protocol to evaluate the produced music. We test (1) audio fidelity (accuracy of the generated music as a perceptually clean and musically plausible outcome) by using audio-language embedding models to test (2) text-music alignment (accuracy of the generated audio with respect to the original emotional prompt), and (3) frequency domain indicators which measure the spectral centroid, spectral roll-off, bandwidth, energy distribution, and other acoustic features related to timbre, brightness, density, and environmental suitability. This evaluation design links high-level affective language to low-level acoustic structure, and is more transparent to the analysis of whether the generated music is indeed affective.

Unlike previous music pipelines that convert text to music directly, our pipeline introduces an explicit step of emotional mapping using an LLM in between the human intention and music generation. This intermediate representation enables the creative process to be more controllable, editable and explainable. It also allows for iterative interaction: a composer may ask for modifications like "make it more fragile", "less sentimental", or "darker in the low-frequency texture", and the system can change the structured parameters before the regeneration. Our work creates a realistic technical solution to foster affective music creation in both artistic and real-world soundscape scenarios by integrating natural language reasoning, parameterized music generation, and frequency-aware evaluation.

## 2. Related Work

### 2.1. Music Generation

Over the past decades, automatic music generation has been explored in a variety of ways, ranging from rule-based composition to symbolic sequence modelling to the latest large-scale neural audio generation. The symbolic representation - such as a sequence of notes, a piano-roll or MIDI events - has dominated early neural approaches, which clearly model pitch, duration, velocity and rhythm. These symbolic systems offer some degree of interpretability of control, yet are restricted in terms of modelling timbre, performance nuances, and realistic audio production.

Recent generative models have shifted toward audio-domain or neural-token-based music synthesis. Jukebox [4] models compressed audio representations with autoregressive transformers and demonstrates the ability to generate coherent music with genre, artist, and lyric conditioning. Subsequent text-conditioned systems further establish natural language as a practical interface for music creation. MusicLM [1] generates high-fidelity music from textual descriptions using hierarchical audio token modeling. MusicGen [3] proposes a simpler and controllable architecture for music generation based on discrete audio tokens. AudioLDM [8] adopts latent diffusion modeling for text-to-audio generation, showing that diffusion-based approaches can synthesize diverse audio content from language prompts.

This creates a high-quality generation but most of the approaches use a raw text prompt directly as a conditioning signal. This makes the connection between the emotional language and the musical structure implicit. Examples: If the description is "tired but hopeful, like walking home in light rain", tempo and dynamics are less, texture is sparse, harmonic tension is moderate, and the color of the music is darker, with a more spectral quality. But in the past, models rarely made such parameter intermediate properties available to the user. This makes the generation process difficult to interpret, revise and control, therefore.

Unlike the previous direct text-to-music pipelines, we introduce a large language model as an affective mapping module. In contrast to directly feeding the original prompt into the music generator, we first encode emotional and metaphorical language into music parameters, such as tempo, dynamics, harmony, density, register, instrumentation tendency, and frequency related targets. This intermediate representation not only increases the controllability of the process but also facilitates an iterative human-AI interaction in music creation.

### 2.2. Quality Evaluation of Synthesized Music

Evaluating synthesized music is challenging because music quality depends on multiple factors, including perceptual fidelity, semantic consistency, emotional expression, musical structure, and acoustic characteristics. Existing works commonly evaluate generated audio using subjective listening tests, objective audio metrics, or representation-based similarity measures. For audio fidelity, metrics such as Fréchet Audio Distance [7] are widely used to estimate the distributional similarity between generated and real audio. These metrics provide useful reference-free evaluation signals, but they do not fully explain whether the generated music matches a specific textual or emotional intention.

With the development of audio-language representation learning, cross-modal alignment has become an important evaluation direction. MuLan [6] learns a joint embedding space between music audio and natural language, enabling text-audio retrieval and semantic similarity measurement. CLAP [5] learns audio concepts from natural language supervision and has been widely adopted for audio-text alignment evaluation. These models allow researchers to estimate whether generated audio is semantically consistent with a given text prompt. However, embedding-based alignment mainly measures high-level semantic correspondence and may overlook low-level acoustic evidence related to musical emotion.

Music emotion and perceived atmosphere are also closely related to frequency-domain properties. Features such as spectral centroid, spectral bandwidth, spectral roll-off, chroma distribution, onset strength, tempo, and energy distribution can reflect brightness, density, rhythmic activity, harmonic color, and timbral tension. Audio analysis tools such as librosa [9] provide reliable implementations for extracting these interpretable features. Emotion-oriented datasets such as DEAM [2] further show that musical affect is often described through continuous dimensions such as valence and arousal, which are influenced by both temporal and spectral acoustic cues.

Hence, it is important to consider more than just fidelity or text-audio similarity within synthesized music for affective music generation. In this paper, music generated under three complementary aspects is analyzed: audio fidelity, text-music alignment and frequency-domain indicators. This evaluation method enables us to evaluate not only the plausibility of the sound produced and congruence with the prompt, but also the acoustic structure in regard to the desired emotional expression.

# 3. Method

## 3.1. Overview

We suggest a framework to create music by mapping emotional language in free form into structured music control and to compare created music against both a semantic and acoustic criteria with the added consideration of affect. Our method aims at producing music that is perceptually plausible, semantically coherent with the text and acoustically similar to the affect the listener is intended to grasp, given a handwritten prompt describing an emotional state, atmosphere or listening context. Unlike the traditional text-to-music pipelines directly passing the original text prompt to the generation model, our framework adds a large language model (LLM) as an intermediate affective mapper. The LLM decodes the emotional and metaphorical information in the prompt and translates it into concrete musical parameters (e.g. tempo, dynamics, harmony, density of the texture, register, instrumentation tendency and spectral targets). These parameters are then input into a more controllable conditioning input to the following music generation model.

The overall difference between direct text-to-music generation and our proposed framework can be formulated as follows. In a conventional pipeline, the generated music is directly conditioned on the raw text prompt:

$$\hat{y} = G(x), \tag{1}$$

where x denotes the original text prompt, G(·) is a text-conditioned music generation model, and $\hat{y}$ is the generated music waveform. Although this formulation is simple, the mapping from emotional language to musical structure is implicit and difficult to inspect.

In contrast, our method decomposes the generation process into an affective mapping stage and a music synthesis stage:

$$z = M_{LLM}(x), \hat{y} = G(z), \tag{2}$$

where $M_{LLM}(\cdot)$ represents the LLM-based affective mapper and z is the structured musical representation inferred from the original prompt. This decomposition reveals the musical intention in the intervening moments prior to synthesis. As a result, the users can be able to view and adjust the musical parameters prior to regeneration, allowing the system to create a more transparent association between human affective language and the generated sound.

## 3.2. Prompt-to-Parameter Transformation

Our framework's key concept is the prompt-to-parameter transformation. Descriptions of music are often subjective, abstract, and metaphorical, as is the human description of it. A prompt like "tired but hopeful, like walking home in light rain" does not contain indications of tempo, key, rhythm, timbre or frequency profile. But it suggests a set of musical tendencies - slow or moderate tempo, soft dynamics, sparse texture, mild harmonic ambiguity, and subdued spectral brightness. The LLM is used to understand, from text, the technical musical information that should be used, rather than the user having to enter this information separately.

Given an input prompt, the LLM produces a structured representation that contains both affective interpretation and generation-oriented musical descriptors. This transformation can be written as:

$$z = \{a, t, d, h, \rho, r, i, s, p\}, \tag{3}$$

where a denotes the high-level affective interpretation, t denotes tempo, d denotes dynamics, h denotes harmonic color and tension, ρ denotes texture density, r denotes register, i denotes instrumentation tendency, s denotes spectral targets, and p denotes the refined generation prompt.

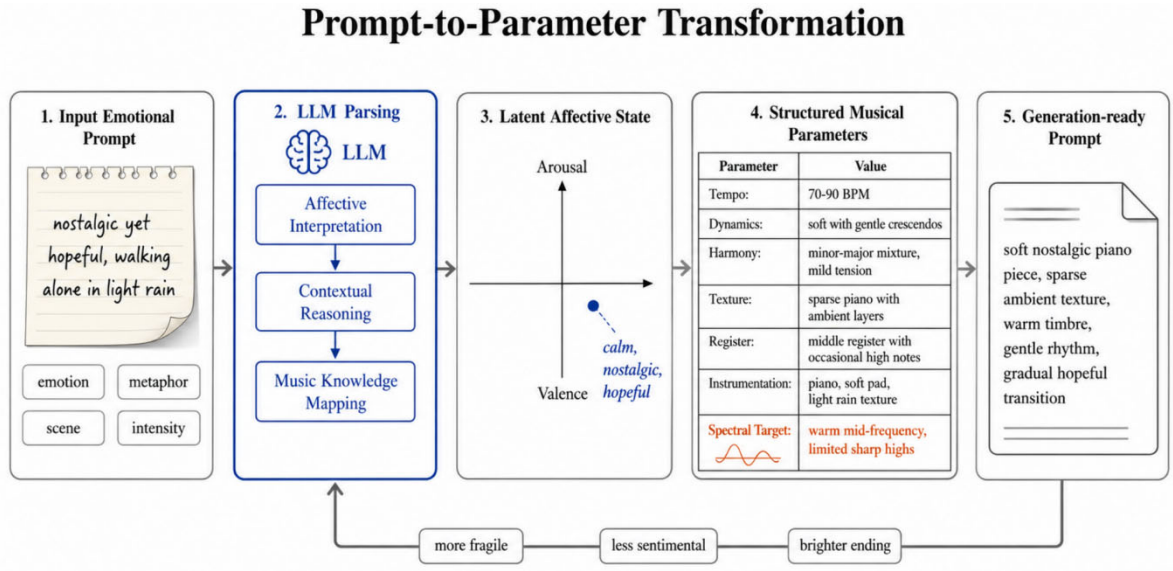


**Figure 2.** Prompt-to-Parameter Transformation.

Compared with the original prompt x, the representation z provides a more explicit and editable description of the intended music.

We adopt a constrained instruction template for the LLM to get stable and controllable outputs. The template encourages the model to explain the emotion in the cue and to describe it musically in terms of specific musical words, not general adjectives. For instance, given the cue "nostalgic, hopeful, walking alone, in light rain after long day," the LLM might result in a slow to moderate tempo, soft dynamics and gentle crescendos, minor major harmonic mixture, sparse piano texture and ambient layers, middle register emphasis, warm mid-frequency energy, and limited high-frequency sharpness. These descriptors are then translated to a more specific generation prompt, for example, a "soft nostalgic piano piece with sparse ambient texture, gentle rhythm, warm timbre and a gradual hopeful transition". This way, the LLM acts as a comprehensible link between intonation and controlled music parameters.

Iterative editing is also possible with this intermediate representation. If a user requests an edit, like "make it more fragile" or "reduce the sentimental feeling", the LLM does not have to start from scratch in order to produce the appropriate revision. Instead, it can modify the parameters associated with the textures, such as lower texture density, softer dynamics, higher instability, or tension in the harmony. Thus, the prompt-to-parameter conversion enhances the controllability and explainability of affective music generation.

### 3.3. Evaluation Protocol

The produced music is judged with regard to three complementary aspects: fidelity, match to the text, and consistency in the frequency domain. This design comes from the premise that good music for affect should meet multiple criteria. It needs to be natural and musically plausible, needs to be congruent with the user's emotional cue and the acoustic shape of the formal needs to facilitate the affective effect. Thus, our evaluation procedure is based on perceptual quality measurement, cross-modal semantic similarity and interpretable spectral analysis.

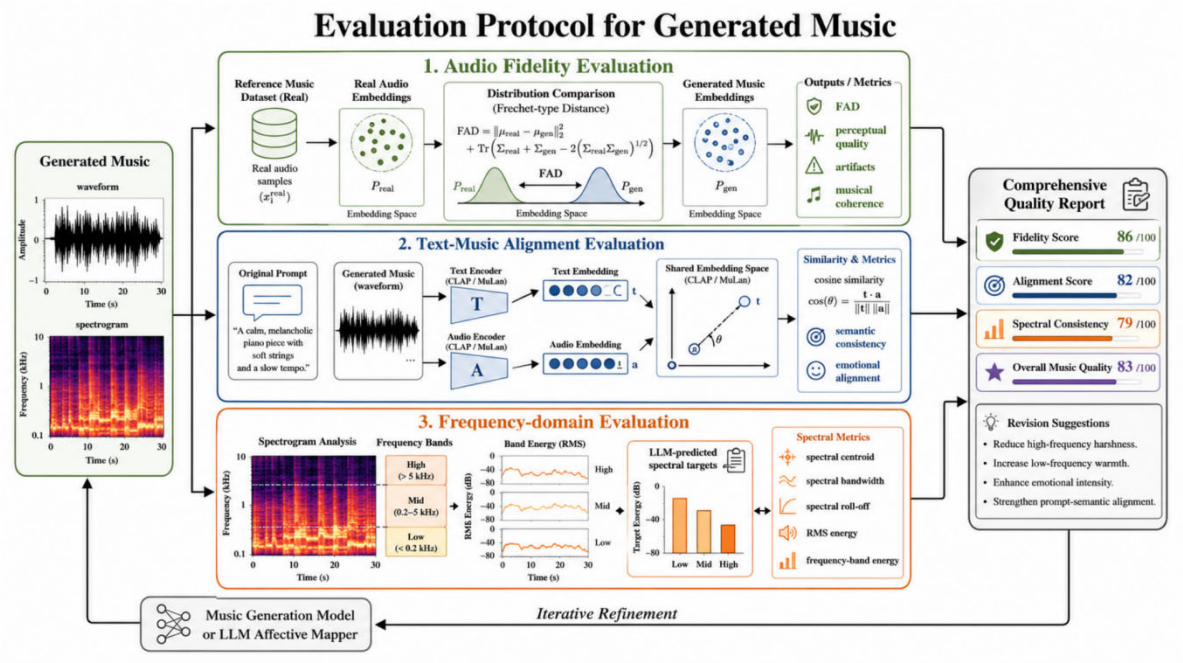


**Figure 3.** Evaluation Protocol.

In terms of audio fidelity, we examine how close the created music is to real music in terms of perceptual quality and acoustic distribution. By building on previous research, we propose a metric for measuring distributional distance between generated music and reference music called Fréchet Audio Distance (FAD). Given audio embeddings extracted from a reference set and a generated set, FAD is computed as:

$$FAD = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}), \tag{4}$$

where $(\mu_r, \Sigma_r)$ and $(\mu_g, \Sigma_g)$ denote the mean and covariance of the reference and generated audio embeddings, respectively. A lower FAD indicates that the generated music distribution is closer to that of real music, suggesting better audio fidelity and fewer perceptual artifacts.

Evaluation for text-music alignment is done by checking the semantic consistency between the output and the input text prompt. To embed the text and the created audio into a common feature space we employ representation models that rely on audio-language, for example, CLAP or MuLan. The alignment score is computed by cosine similarity:

$$S_{align}(x, \hat{y}) = \frac{E_T(x)^\top E_A(\hat{y})}{||E_T(x)||_2 ||E_A(\hat{y})||_2}, \tag{5}$$

where $E_T(\cdot)$ is the text encoder, $E_A(\cdot)$ is the audio encoder, x is the original emotional prompt, and $\hat{y}$ is the generated music. Higher alignment scores mean that the music created better aligns with the semantic and emotional content of the prompt. In practice, this score can be calculated using the original prompt and the LLM-enhanced prompt to see if the LLM-generated music reflects the user's original intention and the structured musical description it created.

To assess the consistency in the frequency domain, we check if the generated music is consistent with the acoustic properties of the inferred LLM spectral targets. In addition to semantic similarity, the frequency domain features are important evidence because the perception of emotion in music is related to timbre, brightness, density and energy distribution. We learn meaningful spectral attributes like spectral centroid, spectral bandwidth, spectral roll-off, chroma distribution, root mean square energy, onset strength and frequency-band energy. Among them, spectral centroid is used to describe the center of mass of the spectrum and is commonly associated with perceived brightness:

$$C_t = \frac{\Sigma_k f_k |X_t(k)|}{\Sigma_k |X_t(k)|}, \tag{6}$$

where $X_t(k)$ denotes the magnitude spectrum at time frame t and frequency bin k, and $f_k$ is the corresponding frequency. The higher the centroid, the brighter or sharper the timbre, and the lower the centroid the darker or warmer the timbre. These features can be used to compare with the LLM-predicted spectral targets, to check whether the generated music is acoustically similar to the desired emotional description.

Overall, our evaluation forms a closed loop from language to parameters, from parameters to audio, and from audio back to semantic and acoustic measurements:

$$x \rightarrow z \rightarrow \hat{y} \rightarrow \{fidelity, alignment, frequency\}. \tag{7}$$

The closed-loop allows us to not only analyse the music that's created from the file, but to analyse the music as the end product of an understandable affective mapping process. It can be used to check the correctness of the musical parameters given by LLM and compare whether the synthesized music matches the user's emotional purpose.

# 4. Experiments

## 4.1. Experimental Setup

We experiment to see if the proposed LLM-guided affective mapping framework can enhance controllability, semantic alignment and acoustic consistency of music generation conditioned on text. We are interested in making perceptually plausible music, but we also want to make sure that the music is still conveying the emotional intention of the input prompt, so there are both semantic and signal-level measurements included.

Four affective music generation benchmarks are built, one of which is small and includes 40 handwritten emotional prompts. The following prompts have been created to reflect a variety of affective conditions and contexts of listening such as nostalgia, loneliness, calmness, anxiety, hopefulness, tension, warmth, and fatigue. There is an emotional description and contextual scene to every prompt as in "a lonely walk through a quiet city at midnight" or "tired but hopeful, like returning home in light rain." Using various techniques, music is created for every prompt and compared by means of audio fidelity, text music alignment and frequency domain indicators.

The target duration is set to 30 seconds for each of the generated samples. The feature is extracted from all the audio resampled to 44.1 kHz. For text-music alignment, we employ the CLAP-based text-audio similarity (TAS); for audio fidelity, we use the Fréchet Audio Distance (FAD); and for audio consistency in the frequency domain, we calculate several spectrally based distances. The frequency domain metrics are spectral centroid, spectral bandwidth, spectral roll-off, RMS energy and low/mid/high frequency energy ratio. Moreover, we perform a small-scale subjective listening test in which participants make ratings for the overall quality, emotional consistency, and listening comfort of the created music.

## 4.2. Music Samples and Prompt Categories

To better analyze the behavior of different systems, we divide the prompts into four affective categories: calm, nostalgic, tense, and hopeful. Each category contains 10 prompts. The calm category emphasizes low arousal and smooth acoustic texture; the nostalgic category often requires warm timbre, sparse texture, and moderate harmonic ambiguity; the tense category contains higher arousal, stronger rhythmic activity, and sharper spectral characteristics; and the hopeful category focuses on gradual emotional transition, brighter harmonic color, and more open register.

**Table 1.** Prompt categories used in our affective music generation benchmark. Each category contains 10 handwritten prompts.

| Category | # Prompts | Representative Emotion | Expected Musical Tendency |
|---|---|---|---|
| Calm | 10 | quiet, relaxed | slow tempo, soft dynamics, low brightness |
| Nostalgic | 10 | warm, reflective | sparse texture, warm timbre, mild tension |
| Tense | 10 | anxious, uneasy | higher density, sharper timbre, stronger rhythm |
| Hopeful | 10 | gentle, positive | gradual crescendo, brighter register, stable harmony |

Table 1 summarizes the prompt categories and representative musical expectations. These expectations are not directly used as ground-truth labels for generation, but serve as analysis references when interpreting the generated results.

## 4.3. Compared Methods

We compare our method with three baseline systems. In Direct Prompt, the handwritten emotional prompt is fed directly into the music generation model without any intermediate processing. The second baseline (Prompt Rewrite) takes an LLM to generate a more fluent

version of the prompt for generation, but without structured musical parameters. The third baseline is Keyword Control, which draws a number of keywords from the input prompt and adds them to the generation prompt, such as "slow", "soft", or "piano". The method developed in our research is different from these baselines as we map emotional language explicitly into structured musical parameters and then use the structured musical parameters to guide the generation process.

All the methods employ the same music generation model on the downstream and the same duration of music generation in order to compare the methods fairly. The difference is only related to the construction of the input condition. This setting enables us to examine the effect of affective mapping through LLM.

## 4.4. Quantitative Results

Table 2 reports the main quantitative comparison. Our method is the most successful on three key evaluation dimensions. Our method achieves improvement in the alignment of CLAP from 0.412 to 0.536 compared to Direct Prompt, thus the produced music is closer to the emotional text. It also lowers the FAD value from 5.87 to 4.31, which means that it has improved audio fidelity. The spectral consistency score rises from 0.641 to 0.792, indicating that the music generated more closely adheres to the frequency domain tendencies as expected.

**Table 2.** Main quantitative comparison. Higher is better for CLAP alignment and spectral consistency, while lower is better for FAD.

| Method | CLAP Alignment ↑ | FAD ↓ | Spectral Consistency ↑ |
|---|---|---|---|
| Direct Prompt | 0.412 | 5.87 | 0.641 |
| Prompt Rewrite | 0.458 | 5.42 | 0.683 |
| Keyword Control | 0.471 | 5.18 | 0.701 |
| Ours | 0.536 | 4.31 | 0.792 |

We further report subjective listening results in Table 3. Each sample is rated on a scale from 1 to 5, with a higher score denoting better performance. Our way achieves the best results in terms of overall quality, emotional consistency and listening comfort. The improvement in emotional consistency is particularly evident, corroborating the fact that statements of emotional cues were explicitly converted to musical parameters prior to generation.

**Table 3.** Subjective listening study results. Scores range from 1 to 5. Higher is better.

| Method | Overall Quality ↑ | Emotional Consistency ↑ | Listening Comfort ↑ |
|---|---|---|---|
| Direct Prompt | 3.42 | 3.18 | 3.51 |
| Prompt Rewrite | 3.61 | 3.46 | 3.64 |
| Keyword Control | 3.73 | 3.58 | 3.70 |
| Ours | 4.16 | 4.08 | 4.21 |

Table 4 shows the CLAP alignment results across different affective categories. We consistently achieve better results than the baselines for calm, nostalgic, tense and hopeful prompts. The gains are particularly large for nostalgic and hopeful prompts, which often contain metaphorical or subtle emotional descriptions. This indicates that the LLM affective mapper can be useful for mapping abstract emotional language to musically meaningful conditions.

**Table 4.** CLAP alignment scores across different affective categories. Higher is better.

| Method | Calm ↑ | Nostalgic ↑ | Tense ↑ | Hopeful ↑ |
|---|---|---|---|---|
| Direct Prompt | 0.426 | 0.398 | 0.431 | 0.392 |
| Prompt Rewrite | 0.469 | 0.441 | 0.472 | 0.451 |
| Keyword Control | 0.481 | 0.456 | 0.487 | 0.459 |
| Ours | 0.548 | 0.527 | 0.541 | 0.529 |

To analyze the contribution of each component, we conduct an ablation study in Table 5. There is a significant decrease in alignment, and spectral consistency, when the structured parameter representation is removed. The frequency domain consistency is influenced by removing the spectral target field, while iterative refinement is influenced by its removal to not only subjective quality but also alignment. As observed from these results, the individual components can each make a contribution to the overall performance and the whole framework gives the best balanced result.

**Table 5.** Ablation study of the proposed framework. Higher is better for CLAP alignment and spectral consistency, while lower is better for FAD.

| Variant | CLAP Alignment ↑ | FAD ↓ | Spectral Consistency ↑ |
|---|---|---|---|
| Ours w/o Structured Parameters | 0.487 | 4.89 | 0.711 |
| Ours w/o Spectral Targets | 0.511 | 4.58 | 0.724 |
| Ours w/o Iterative Refinement | 0.503 | 4.67 | 0.746 |
| Ours Full | 0.536 | 4.31 | 0.792 |

## 4.5. Qualitative Results

Figure 4 shows representative qualitative examples from different prompt categories. Our framework generates music that more accurately reflects the desired emotional evolution of music, when compared against the baseline methods. Our method is generally more effective for calm prompts, with smoother dynamics and lower spectral brightness. Music generated for nostalgic prompts has less texture and warmer mid-frequency energy. Tense prompts: the model is more dense in the rhythmic domain, more high frequency content and also maintains musical coherence. Our approach to hopeful prompts often includes building elements in the top register and/or building to a crescendo for the emotional transfer.

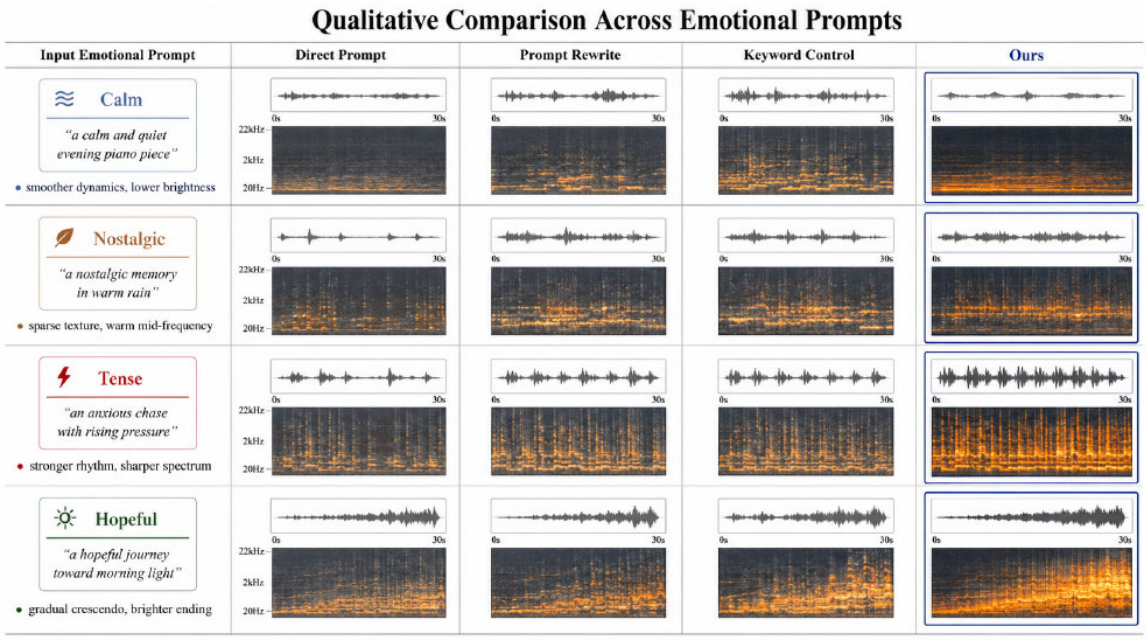


**Figure 4.** Qualitative comparison of generated music under different emotional prompts. Our method better captures affective intention by explicitly mapping the prompt into structured musical parameters before generation.

Figure 5 visualizes the frequency-domain analysis of generated music. The baseline from the direct prompt shows less regular spectral characteristics and frequently has high frequency noise. Our approach, on the other hand, generates spectrogram patterns more aligned with the spectral targets predicted by the LLM. As such, calm and nostalgic prompts display a softer high frequency sharpness and a smoother energy distribution in the mid frequencies than do tense prompts. All of these observations agree with the quantitative results of better spectral consistency.

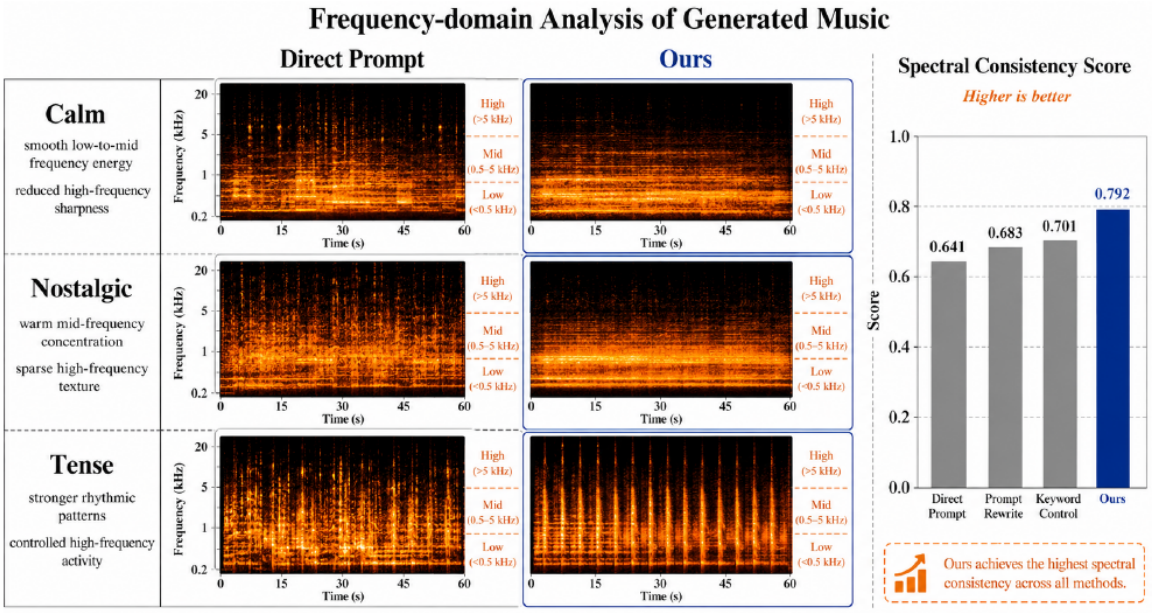


**Figure 5.** Frequency-domain visualization of generated music.

### 4.6. Result Analysis

The experimental results show that explicit affective mapping is beneficial for text-conditioned music generation. Emotional prompts can directly generate plausible music, although the generation model may not be able to capture subtle and metaphorical descriptions. Prompt rewriting leads to increased fluency without a clear musical control. Keyword control does not offer much guidance and is not able to express complex relationships among emotion, musical structure and acoustic properties.

In contrast, our approach has proposed an intermediate representation with structure which can decompose the affective intention into controllable musical factors. This enhances the semantic alignment as the generation prompt is more musically specific. It also has the advantage of increasing the frequency domain consistency as the LLM explicitly predicts the tendency of the frequency domain which can be compared to the synthesised audio. Ablation results also demonstrate that both structured parameters and spectral targets, as well as iterative refinement, play an important role in the final performance. The proposed framework, as a whole, forms more interpretable and controllable solution for affective music generation.

## 5. Conclusion

In this paper, we introduced an LLM guided affective music creation framework that connects free-form emotional language to controllable music creation by an explicit prompt to parameter transformation. Our method uses a large language model to infer structured musical descriptors such as tempo, dynamics, harmony, texture, register, instrumentation and spectral targets, instead of using raw text prompts as black-box conditions, making the generation process more interpretable and editable. We also proposed a multi-dimensional evaluation protocol which evaluates audio fidelity, text-music alignment, and frequency-domain consistency. Our results demonstrate that our framework achieves better performance than direct prompt generation and other prompt-processing baselines, with 0.536 CLAP alignment, 4.31 FAD, and 0.792 spectral consistency compared to 0.412, 5.87, and 0.641 respectively. The results indicate that LLM-based affective mapping can be an effective approach to create music which is more representative of human emotion intention and acoustically consistent.

## References


[1] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Frank, C., & Zeghidour, N. (2023). MusicLM: Generating music from text. arXiv preprint arXiv:2301.11325. https://doi.org/10.48550/arXiv.2301.11325.

[2] Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. PLOS ONE, 12(3), e0173392. https://doi.org/10.1371/journal.pone.0173392

[3] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple and controllable music generation. Advances in Neural Information Processing Systems, 36.

[4] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341. https://doi.org/10.48550/arXiv.2005.00341

[5] Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023). CLAP: Learning audio concepts from natural language supervision. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1–5. https://doi.org/10.1109/ICASSP40776.2023.10096654

[6] Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., & Ellis, D. P. W. (2022). MuLan: A joint embedding of music audio and natural language. Proceedings of the International Society for Music Information Retrieval Conference.

[7] Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. Proceedings of Interspeech, 2350–2354. https://doi.org/10.21437/Interspeech.2019-2680

[8] Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., & Plumbley, M. D. (2023). AudioLDM: Text-to-audio generation with latent diffusion models. Proceedings of the International Conference on Machine Learning.

[9] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. Proceedings of the Python in Science Conference, 18–25. https://doi.org/10.25080/Majora-7b98e3ed-003.