A Attribute Information

Table 8 shows the detailed pre-defined musical attribute values. The value *NA* of each attribute refers to that this attribute is not mentioned in the text. Objective attributes can be extracted from MIDI files with heuristic algorithms and subjective attributes are collected from existing datasets, as shown in Table 9.

Table 8: Detailed attribute values.

Attributes	Values
Instrument	28 instruments: piano, keyboard, percussion, organ, guitar, bass, violin, viola, cello, harp, strings, voice, trumpet, trombone, tuba, horn, brass, sax, oboe, bassoon, clarinet, piccolo, flute, pipe, synthesizer, ethnic instrument, sound effect, drum. Each instrument: 0: played, 1: not played, 2: NA
Pitch Range	0-11: octaves,12: NA
Rhythm Danceability	0: danceable, 1: not danceable, 2: NA
Rhythm Intensity	0: serene, 1: moderate, 2: intense, 3: NA
Bar	0: 1-4 bars, 1: 5-8 bars, 2: 9-12 bars, 3: 13-16 bars, 4: NA
Time Signature	0: 4/4, 1: 2/4, 2: 3/4, 3: 1/4, 4: 6/8, 5: 3/8, 6: other tempos, 7: NA
Key	0: major, 1: minor, 2: NA
Tempo	0: slow (<=76 BPM), 1: moderato (76-120 BPM), 2: fast (>=120 BPM), 3: NA
Time	0: 0-15s, 1: 15-30s, 2: 30-45s, 3: 45-60s, 4: >60s, 5: NA
Artist	0-16 artists: Beethoven, Mozart, Chopin, Schubert, Schumann,J.S.Bach, Haydn, Brahms, Handel, Tchaikovsky, Mendelssohn,Dvorak, Liszt, Stravinsky, Mahler,Prokofiev, Shostakovich, 17: NA
Genre	21 genres: new age, electronic, rap, religious,international,easy listening, avant garde, RNB, latin, children, jazz, classical, comedy, pop, reggae, stage, folk, blues, vocal, holiday, country, symphony Each genre: 0: with, 1: without, 2: NA
Emotion	0-3: the 1-4 quadrant in Russell's valence-arousal emotion space 4: NA

Table 9: Extraction methods and sources of each attributes.

Type	Attribute	Extraction Method
Objective	Instrument	directly extracted from MIDI
	Pitch range	calculated based on the pitch range
	Rhythm danceability	judged with the ratio of downbeat
	Rhythm intensity	judged with the average note density
	Bar	directly extracted from MIDI
	Time signature	directly extracted from MIDI
	Key	judged with the note pitches based on musical rules
	Tempo	directly extracted from MIDI
	Time	derived from the time signature and the number of bars
Subjective	Artist	provided by a classical music dataset in MMD [36]
	Genre	provided by MAGD[13], a classical music dataset in MMD [36] and Symphony [39]
	Emotion	provided by EMOPIA [16] and the emotion-gen dataset

B Experiments

B.1 User study with baselines

In the user study, participants were provided with generated music samples along with their corresponding textual prompts. For each text description, each model (i.e., BART-base, GPT-4, MuseLM) generated three different music clips. In each questionnaire, three samples generated with the same

text conditions were randomly picked from samples generated by BART-base, GPT-4, and MuseLM respectively as a group. Each participant was asked to evaluate 7 groups for comparison. Three subjective metrics, musicality, controllability, and an overall score, are rated on a scale of 1 (lowest) to 5 (highest). The participants were first requested to evaluate their music profession level, as depicted in Table 10. To ensure the reliability of the assessment, only individuals with at least music profession level 3 were selected, resulting in a total of 19 participants. Secondly, they were instructed to independently evaluate two separate metrics: musicality and controllability, ensuring that scoring for one metric did not influence the other. They are also asked to give an overall score to evaluate the generated music comprehensively. For the collected results, we computed the mean and variance for each metric. The results can be found in Table 4.

Table 10: Music Profession Level

Level	Description
1	I rarely listen to music.
2	I haven't received formal training in playing or music theory, but I often listen to music and have my preferred styles, musicians, and genres.
3	I have some basic knowledge of playing an instrument or music theory, but I haven't received formal training.
4	I haven't received formal training, but I have self-taught myself some aspects such as music theory or playing an instrument. I am at an amateur level (e.g., CCOM piano level 6 or above).
5	I have received professional training in a systematic manner.

B.2 Objective Comparison with baselines

In this section, we introduce how to calculate the objective metric, the average sample-wise accuracy (ASA), in Table 4. As for MuseLM, ten music clips are generated per prompt and we report ASA of them among the overall standard test set. Since it is labor-intensive to leverage GPT-4 with the official web page, we only guide GPT-4 to produce five music clips per prompt and calculate the ASA of 21 prompts randomly sampled from the standard test set. Besides, we utilize the released text-tune BART-base checkpoint[14] to generate five music clips per prompt and report the ASA of 44 prompts randomly chosen from the standard test set.

B.3 Text-to-attribute understanding

As shown in Table 11, all attribute control accuracy is close or equal to 100%, which indicates our model with multiple classification heads in the text-to-attribute understanding stage performs quite well.

B.4 Details of Analysis on Attribute-to-music Generation

Attribute Control Accuracy We report the control accuracy for each attribute on the test dataset, as shown in Table 12. The average attribute control accuracy of 80.42%, which provides substantial evidence for the model's proficiency in effectively controlling music generation using music attributes.

Study on Control Methods To verify the effectiveness of the control method in the attribute-to-music generations stage, we compare *Prefix Control* with two methods: *Embedding* and *Conditional LayerNorm*. For efficiency, we conducted this study on reduced-size models as follows: The backbone model of this experiment is a 6-layer Linear Transformer with causal attention. The hidden size is 512 and the FFN hidden size is 2048. The other experiment configuration is the same as Section 4.1. Since the control accuracy of objective attributes can be easily calculated, we only need to measure the controllability of each subjective attribute in listening tests. The control accuracy of each attribute is shown in Table 12. Finally, the average attribute control accuracy can be calculated based on the accuracy results from both types of attributes. To measure the controllability of subjective attributes

[14] `https://huggingface.co/sander-wood/text-to-music`

Table 11: Attribute control accuracy (%) of the text-to-attribute understanding model. I: Instrument.

Attribute	Accuracy(%)	Attribute	Accuracy(%)	Attribute	Accuracy(%)
I_piano	100.00	I_clarinet	99.92	Genre_comedy_spoken	100.00
I_keyboard	99.92	I_piccolo	99.94	Genre_pop_rock	100.00
I_percussion	100.00	I_flute	99.62	Genre_reggae	100.00
I_organ	100.00	I_pipe	100.00	Genre_stage	100.00
I_guitar	99.92	I_synthesizer	100.00	Genre_folk	100.00
I_bass	99.84	I_ethnic_instruments	99.98	Genre_blues	100.00
I_violin	99.92	I_sound_effects	99.98	Genre_vocal	100.00
I_viola	99.96	I_drum	100.00	Genre_holiday	100.00
I_cello	99.92	Genre_new_age	99.98	Genre_country	100.00
I_harp	100.00	Genre_electronic	100.00	Genre_symphony	100.00
I_strings	99.96	Genre_rap	100.00	Bar	100.00
I_voice	99.70	Genre_religious	100.00	Time Signature	100.00
I_trumpet	99.96	Genre_international	100.00	Key	100.00
I_trombone	99.94	Genre_easy_listening	100.00	Tempo	99.84
I_tuba	100.00	Genre_avant_garde	100.00	Octave	100.00
I_horn	99.94	Genre_rnb	100.00	Emotion	99.80
I_brass	100.00	Genre_latin	100.00	Time	100.00
I_sax	99.84	Genre_children	100.00	Rhythm Danceability	100.00
I_oboe	99.94	Genre_jazz	100.00	Rhythm Intensity	99.88
I_bassoon	99.96	Genre_classical	100.00	Artist	100.00

Table 12: Accuracy (%) of each attribute for attribute-to-music generation. I: Instrument.

Attribute	Accuracy(%)	Attribute	Accuracy(%)
I_piano	96.20	I_clarinet	90.63
I_keyboard	79.55	I_piccolo	86.67
I_percussion	65.19	I_flute	86.73
I_organ	80.55	I_pipe	70.73
I_guitar	91.81	I_synthesizer	78.28
I_bass	93.11	I_ethnic_instruments	77.69
I_violin	87.88	I_sound_effects	51.74
I_viola	92.03	I_drum	95.96
I_cello	86.50	Bar	71.80
I_harp	74.87	Time Signature	99.14
I_strings	86.08	Key	57.42
I_voice	75.82	Tempo	92.71
I_trumpet	84.86	Octave	61.56
I_trombone	84.64	Time	65.82
I_tuba	93.08	Rhythm Danceability	88.04
I_horn	80.13	Rhythm Intensity	80.47
I_brass	77.27	Genre	73.08
I_sax	81.74	Emotion	69.45
I_oboe	85.23	Artist	50.03
I_bassoon	90.72		

Prompt

Please combine the following sentences to one paragraph:
The music is imbued with [EMOTION]. The [KEY] key in this music provides a powerful and memorable sound. The song progresses through [NUM_BARS] bars. [TIME_SIGNATURE] is the meter of the music.

ChatGPT Response

The music is imbued with [EMOTION], creating a captivating and emotionally rich experience. It is composed in the [KEY] key, which adds a powerful and memorable sound to the composition. The song unfolds through [NUM_BARS] bars, with [TIME_SIGNATURE] serving as the meter, further enhancing its rhythmic and melodic structure.

Figure 3: ChatGPT refines concatenated templates in the prompt.

Prompt

I want a music in the ABC notation format and assign proper MIDI instrument IDs as the following:

X:1
T:Forest Stroll
M:3/4
L:1/4
Q:1/4=80
K:C
V:1
%%MIDI program 90 % Synthesizer
E/2F/2|G2A|G2F|E2D|C2E|F2G|A3-|A2E/2F/2|
G2A|G2F|E2D|C2E|F2G|A2F|E2D|C3||
V:2
%%MIDI program 68 % Oboe
G/2A/2|B2C|B2A|G2F|E2G|A2B|c3-|c2G/2A/2|
B2C|B2A|G2F|E2G|A2B|c2A|G2F|E3||
V:3
%%MIDI program 73 % Flute
z/2A/2|B2C|B2A|G2F|E2G|A2B|c3-|c2z/2A/2|
B2C|B2A|G2F|E2G|A2B|c2A|G2F|E3||
V:4
%%MIDI program 42 % Cello
C/2D/2|E2F|E2D|C2B,|A,2C|D2E|F3-|F2C/2D/2|
E2F|E2D|C2B,|A,2C|D2E|F2D|C2B,|A,3||

Compose ABC notation music as the format above according to the music description as follow: The music should feature the cello and flute. The music is in 4/4 meter. The song is 31 ~ 45 seconds in length.

GPT-4 Response

X:1
T:Whispering Winds
M:4/4
L:1/4
Q:1/4=80
K:C
V:1
%%MIDI program 73 % Flute
A/2B/2|c2d|e2c|A2G|F2A|B2c|d3-|d2A/2B/2|
c2d|e2c|A2G|F2A|B2c|d2A|G2F|A3||
V:2
%%MIDI program 42 % Cello
C/2E/2|F2G|A2F|C2B,|A,2C|E2F|G3-|G2C/2E/2|
F2G|A2F|C2B,|A,2C|E2F|G2C|B,2A,|C3||

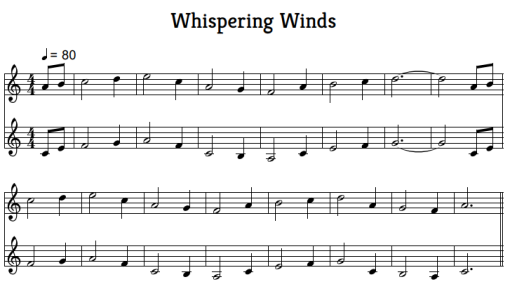

Figure 4: GPT-4 generates ABC notation tunes based on the prompt.

(such as emotion and genre), we invite 12 participants to conduct a listening test. Each participant was provided with 18 music pieces (6 pieces per control method) with corresponding subjective attributes. We asked each participant to answer: 1) Musicality(five-point scale): How similar it sounds to the music composed by a human. 2) Controllability: Does it align with the given attributes. Then we report the musicality and average attribute accuracy score in Table Table 7. The experimental results clearly demonstrate that *Prefix Control* outperforms the other two methods in terms of musicality and controllability.

B.5 Usage of GPT models

Refine texts with ChatGPT As shown in Figure 3, in order to make text descriptions more coherent and fluent, we feed concatenated templates into ChatGPT with a prompt *Please combine the following sentences to one paragraph* and then ChatGPT will give a response containing all templates within a compact paragraph.

Generate ABC notation music with GPT-4 To use the GPT-4 as the baseline method for comparison, we design the instruction to guide GPT-4 as shown in Figure 4. Gpt-4 can only generate symbolic music in ABC notation, so we need to explicitly point out the format. Besides, since GPT-4 can generate various ABC notation formats, some of which cannot be processed by music21, we provide an ABC notation example, teaching GPT-4 to follow its format. Meanwhile, we use the prompt, *Compose ABC notation music as the format above according to the music description as follows: [text descriptions]* to let GPT-4 generate music according to the text description and convert the ABC notations into MIDI for a fair comparison.

C Limitation

This work is mainly about generating symbolic music from text descriptions, which does not consider long sequence modeling especially. To address this, we can employ Museformer [22] as the backbone model, which proposes fine- and coarse-grained attention for handling long sequences.

The attribute set provided in this work represents only a subset of all music attributes. We aim to further explore additional attributes to offer a wider range of control options for music generation, ensuring greater diversity in the creative process.

The possibility of regenerating music based on additional text descriptions to assist users in refining their compositions is an aspect that is worth exploring.