HW2: Controllable Text-to-Music Generation

Student ID: R13921098

1 Project Structure

```
HW2/
 main_pipeline.py              # Main pipeline orchestration
 src/
    retrieval/
       audio_encoder.py      # Stable Audio VAE encoder
       similarity.py         # Cosine similarity retrieval
    captioning/
       audio_captioner.py    # LP-MusicCaps model
    extraction/
       melody_extractor.py   # Chromagram extraction
       rhythm_extractor.py   # Beat/onset detection
       dynamics_extractor.py # Dynamics features
    generation/
       music_generator.py    # MusicGen-Melody, JASCO
    evaluation/
        clap_similarity.py    # CLAP evaluation
        audiobox_aesthetics.py# Aesthetics metrics
        melody_accuracy.py    # Melody comparison (DTW)
 outputs/
    generated/                # Generated audio files
       jasco/                # JASCO outputs
    features/                 # Extracted features
       melody/
       rhythm/
       dynamics/
    results/                  # Evaluation results
 requirements.txt
```

2 Source Code

Repository URL: [https://github.com/PoHsuanLai/DeepMIR-HW2.git]

3 Implementation Details

3.1 Model: JASCO (Best Configuration)

Model: facebook/jasco-chords-drums-melody-1B (1B parameters)

Text Input: Qwen2-Audio-7B-Instruct (Qwen/Qwen2-Audio-7B-Instruct) generates natural language captions focusing on melodic character, rhythm, and harmonic progression without mentioning specific instruments.

Time-Varying Conditions:

- **Melody:** JASCO melody salience matrix (53 MIDI note bins, 36-88) extracted using basic_pitch with threshold 0.3 for sparse, salient pitch activations.
- **Chords:** Extracted from full mix via librosa chroma_stft with template matching (24 chord types: major/minor). Format: (chord_label, timestamp) tuples.
- **Drums:** Separated drum track via Demucs htdemucs model from full mix.

Generation Parameters (Best from CFG Experiments):

- Duration: 10 seconds (fixed)
- Sample rate: 32kHz
- CFG coefficients: $\gamma_{\text{all}} = 3.0$, $\gamma_{\text{txt}} = 3.0$ (equal weighting)
- Configuration: text_focused (best balance between target matching and text adherence)

4 Aggregate Evaluation Results

Note: Results shown are aggregate metrics from the best configuration (text_focused: $\gamma_{\text{all}} = 3.0$, $\gamma_{\text{txt}} = 3.0$) across 9 diverse target files including Western classical (rock, jazz, country) and Chinese traditional music (bamboo flute, piano).

4.1 Overall Performance

- **CLAP Gen↔Target:** 0.4292 (target audio similarity)
- **CLAP Text↔Gen:** 0.3035 (text adherence)
- **CLAP Text↔Target:** 0.2421 (caption quality baseline)
- **Melody Accuracy:** 0.3346 (pitch contour and chroma similarity)
- **Aesthetics CE:** 0.978 (content enjoyment)
- **Aesthetics PQ:** 1.000 (production quality)

5 Sample Evaluation Results

5.1 6_rock_102_beat_3-4.wav

Generated File: `6_rock_102_beat_3-4_generated.wav`

Text Input (Qwen2-Audio Natural Language Caption):

The music has a bouncy and lively feel with a melody that flows smoothly. It maintains a consistent rhythm and beat, moving along at a moderate tempo. There's a sense of brightness to the harmonic progression, giving it an uplifting quality. Overall, the music exudes energy and a vibrant atmosphere.

CLAP Similarity:

- Target ↔ Text: -0.0296
- Text ↔ Generated: 0.1713
- Generated ↔ Target: 0.6201

Meta Audiobox Aesthetics:

- Target: CE=0.555, CU=0.902, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.7000
- Chroma Accuracy: 0.2227
- Pitch Contour Similarity: 0.3877
- Overall: 0.4154

5.2 10_country_114_beat_4-4.wav

Generated File: `10_country_114_beat_4-4_generated.wav`

Text Input (Qwen2-Audio Natural Language Caption):

The melody is bouncy and lively, with a steady beat that gives it a groovy feel. It's not too high-pitched, making it easy to listen to. The harmony isn't too complex, but it contributes to an uplifting and fun atmosphere. Overall, the song has a cheerful mood and energy level that makes you want to move along with the beat.

CLAP Similarity:

- Target ↔ Text: -0.0641
- Text ↔ Generated: 0.2157
- Generated ↔ Target: 0.4463

Meta Audiobox Aesthetics:

- Target: CE=0.562, CU=0.631, PC=1.000, PQ=1.000
- Generated: CE=0.803, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.6887
- Chroma Accuracy: 0.1021
- Pitch Contour Similarity: 0.5430
- Overall: 0.4103

5.3 4_jazz_120_beat_3-4.wav

Generated File: `4_jazz_120_beat_3-4_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] This song features percussion being played at a fast tempo. The beat is of moderate difficulty. The kick is played on the first count of each bar. The hi-hat is played at different parts of the song. There are no voices in this song. This song can be played in a drum instruction video. [10s-20s] This song features percussion being played at a fast tempo. The beat is of moderate difficulty. The kick is played on the first count of each bar. The hi-hat is played at different parts of the song. There are no voices in this song. This song can be played in a club. [20s-30s] This song features percussion being played at a fast tempo. The beat is of moderate difficulty. The kick is played on the first count of each bar. The hi-hat is played at different parts of the song. There are no voices in this song. This song can be played in a drum instruction video. [30s-40s] The low quality recording features a cowbell percussion. The recording is noisy and in mono. [40s-50s] The low quality recording features a cowbell percussion. The recording is noisy and in mono. [50s-60s] This music is instrumental. The tempo is medium with a cowbell percussion. There is no voice in this clip. It is an instrumental clip. There are no other instruments in this song.

CLAP Similarity:

- Target $\leftrightarrow$ Text: 0.3366
- Text $\leftrightarrow$ Generated: 0.4263
- Generated $\leftrightarrow$ Target: 0.5431

Meta Audiobox Aesthetics:

- Target: CE=0.693, CU=0.719, PC=1.000, PQ=1.000
- Generated: CE=0.769, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.6499

- Chroma Accuracy: 0.0487
- Pitch Contour Similarity: 0.4461
- Overall: 0.3483

5.4 - (Bamboo flute)

Generated File: `-D_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] The low quality recording features a mellow harp melody playing. It sounds soft, mellow and the recording is noisy and in mono. [10s-20s] The low quality recording features a mellow piano melody playing over sustained strings melody. It sounds emotional and passionate. [20s-30s] The low quality recording features a mellow electric guitar melody. It sounds sad, emotional and passionate. [30s-40s] This music is a mellow, dulcet instrumental. The tempo is slow with a beautiful violin harmony, piano accompaniment and steady bass line. The music is soft, pensive, melancholic, sentimental, wistful, panned to the right side of the speakers. This music is an exquisite middle eastern instrumental. [40s-50s] This is a classical music piece. It is an instrumental piece. The main melody is being played by a harpsichord. The atmosphere is lively. This piece could be used in the soundtrack of a historical drama TV series during the scenes where the characters are taking a stroll through the meadows. [50s-60s] The low quality recording features a live performance of a pop song that consists of an arpeggiated electric guitar melody, followed by synth pad chords. It sounds passionate, emotional and the recording is noisy.

CLAP Similarity:

- Target $\leftrightarrow$ Text: 0.3647
- Text $\leftrightarrow$ Generated: 0.3751
- Generated $\leftrightarrow$ Target: 0.2051

Meta Audiobox Aesthetics:

- Target: CE=1.000, CU=0.919, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.4005
- Chroma Accuracy: 0.1323
- Pitch Contour Similarity: 0.6220
- Overall: 0.3597

5.5 Hedwig's theme x dizi (Harry Potter)

Generated File: `Hedwig's theme x dizi_60s_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] The low quality recording features a breathy flute melody played over mellow strings melody. It sounds soulful and passionate. The recording is noisy and slightly distorted. [10s-20s] The low quality recording features a breathy flute melody played over mellow strings melody. It sounds soulful and passionate. The recording is noisy and slightly distorted. [20s-30s] This song is played on the bamboo flute. The melody is simple and traditional. The sound is tremulous. There are no voices in this song. This song can be played in a children's kung fu animation movie. [30s-40s] This music is instrumental. The tempo is fast with an enthusiastic flute harmony and bassoon accompaniment. The music is upbeat, catchy, engaging, vivacious, melodic, cheerful, happy and pleasant. This music is an upbeat classical instrumental. [40s-50s] The low quality recording features a breathy flute melody played over mellow keys chords. It sounds soulful and passionate. The recording is noisy and in mono. [50s-60s] This music is instrumental. The tempo is fast with an enthusiastic flute harmony and bassoon accompaniment. The music is upbeat, catchy, engaging, vivacious, melodic, cheerful, happy and pleasant. This music is an upbeat classical instrumental.

CLAP Similarity:

- Target ↔ Text: 0.5191
- Text ↔ Generated: 0.3589
- Generated ↔ Target: 0.6711

Meta Audiobox Aesthetics:

- Target: CE=1.000, CU=1.000, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.1993
- Chroma Accuracy: 0.1253
- Pitch Contour Similarity: 0.5552
- Overall: 0.2765

5.6 (Cover)

Generated File: `_cover_60s_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] This music is instrumental. The tempo is slow with a beautiful Harp melody. The music is dreamy, cascading, ethereal, soothing, calming and relaxing. This music is a Western Classical Harp Solo. [10s-20s] This song contains someone

playing a melody on a harp. This song may be playing live in a concert. [20s-30s] The song is an instrumental. The tempo is medium with a cello playing the lead melody with no other instrumentation. The song is emotional and romantic. The audio quality is poor. [30s-40s] This is a live performance of a quartet. The quartet consists of upright bass, cello, acoustic guitar and violin. The instrumental has a pop feel to it, with the flute playing the melody. [40s-50s] The low quality recording features a flat male vocal talking, after which an accordion melody is playing. It sounds emotional and passionate. The recording is noisy and in mono. [50s-60s] The low quality recording features a live performance of a folk song and it consists of harmonica solo melody played over breathy flute melody. It sounds emotional and passionate. The recording is noisy.

CLAP Similarity:

- Target ↔ Text: 0.3982
- Text ↔ Generated: 0.5211
- Generated ↔ Target: 0.4461

Meta Audiobox Aesthetics:

- Target: CE=1.000, CU=0.938, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.3105
- Chroma Accuracy: 0.0580
- Pitch Contour Similarity: 0.4302
- Overall: 0.2454

5.7 Spirited Away OST (Piano)

Generated File: `Spirited Away OST_60s_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] This is a piano cover of a glam metal music piece. The piece is being played gently on a keyboard with a grand piano sound. There is a calming, relaxing atmosphere in this piece. It could be playing in the background at a coffee shop. [10s-20s] This is a piece that would be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin. [20s-30s] This is a piece that would be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin. [30s-40s] This is a cover of a glam metal music piece. The piece is being played gently on a keyboard with a grand piano sound. There is a calming, relaxing atmosphere in this piece. It could be playing in the background at a coffee shop. [40s-50s] This is a piece that would

be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin. [50s-60s] The low quality recording features a reverberant groovy piano melody. It sounds emotional and passionate.

CLAP Similarity:

- Target ↔ Text: 0.5540
- Text ↔ Generated: 0.5207
- Generated ↔ Target: 0.5332

Meta Audiobox Aesthetics:

- Target: CE=1.000, CU=0.692, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=0.755, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.3047
- Chroma Accuracy: 0.0371
- Pitch Contour Similarity: 0.5344
- Overall: 0.2666

5.8 Mussorgsky: Pictures at an Exhibition (Piano)

Generated File: `Mussorgsky_60s_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] This is a piece that would be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin. [10s-20s] This is a piece that would be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin. [20s-30s] This is a piece that would be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin. [30s-40s] This is a piano cover of a glam metal music piece. The piece is being played gently on a keyboard with a grand piano sound. There is a calming, relaxing atmosphere in this piece. It could be playing in the background at a coffee shop. [40s-50s] This audio contains someone playing a modern piece of music on an acoustic piano. This song may be playing live in a bar with a piano. [50s-60s] This is a piece that would be suitable as calming study music or music for sleeping. It features a relaxing and soothing motif on the piano, being backed by a distant, high pitched and sustained violin.

CLAP Similarity:

- Target ↔ Text: 0.3673

- Text ↔ Generated: 0.4095
- Generated ↔ Target: 0.3169

Meta Audiobox Aesthetics:

- Target: CE=1.000, CU=0.689, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.2844
- Chroma Accuracy: 0.0603
- Pitch Contour Similarity: 0.5713
- Overall: 0.2808

5.9 IRIS OUT (Piano)

Generated File: `IRIS OUT_generated.wav`

Text Input (LP-MusicCaps Caption):

[0s-10s] The low quality recording features a groovy piano melody, groovy bass guitar, shimmering hi hats, punchy kick and snare hits. It sounds energetic and the recording is noisy and in mono. [10s-20s] This is a live performance of a gospel music piece. There is a male vocalist singing melodically in the lead. The piano is playing the melody while the bass guitar is playing in the background. The rhythm is being played by the acoustic drums. The atmosphere is joyful. This piece could be used in the soundtrack of a Christmas movie. [20s-30s] The low quality recording features a live performance of a rock song and it consists of groovy bass guitar, electric guitar melody, shimmering hi hats, punchy kick and snare hits. There are some crowd cheering noises. It sounds groovy, fun and the recording is noisy and in mono. [30s-40s] This is a live performance of a gospel music piece. It is being performed by an orchestra. The main melody is being played by the electric guitar while the bass guitar is playing in the background. The atmosphere is vibrant. This piece could be used in the soundtrack of a Christmas movie. [40s-50s] This is a live performance of a gospel music piece. It is an instrumental piece. The main melody is being played by the piano while the bass guitar is playing in the background. There is a groovy atmosphere to this piece. This piece could be used in the soundtrack of a Christmas movie. [50s-60s] This is a live performance of a folk rock music piece. There is a male vocalist singing melodically in the lead. The melodic background consists of the electric guitar and the bass guitar playing a simple tune. The atmosphere is trippy. This piece could be used in the soundtrack of a comedy movie during the scenes where a character is reminiscing about the good memories.

CLAP Similarity:

- Target ↔ Text: 0.2823

- Text ↔ Generated: 0.3718
- Generated ↔ Target: 0.1138

Meta Audiobox Aesthetics:

- Target: CE=1.000, CU=1.000, PC=1.000, PQ=1.000
- Generated: CE=1.000, CU=1.000, PC=1.000, PQ=1.000

Melody Accuracy:

- Chroma Similarity: 0.2751
- Chroma Accuracy: 0.0046
- Pitch Contour Similarity: 0.3747
- Overall: 0.1968

6 CFG Experiments

6.1 JASCO Configuration Results

Table 1: JASCO CFG Configuration Results

Config	γ_{all}	γ_{txt}	Gen↔Target	Text↔Gen	Melody Acc	Aesthetics CE
text_focused	3.0	3.0	0.3800	0.3339	0.3318	0.915
text_heavy	2.0	4.0	0.3751	0.3163	0.3423	0.949
melody_focused	5.0	1.0	0.3666	0.2517	0.3349	0.942
balanced	4.0	2.0	0.3621	0.2316	0.3395	0.983
melody_heavy	7.0	0.5	0.3218	0.2471	0.3445	0.979

6.2 MusicGen-Melody Configuration Results

Table 2: MusicGen-Melody Guidance Scale Results

Config	Guidance Scale	Gen↔Target	Text↔Gen	Melody Acc	Aesthetics CE
guidance_medium	3.0	0.2146	0.2760	0.2674	1.000
guidance_low	2.0	0.2063	0.2005	0.2625	1.000
guidance_high	5.0	0.1803	0.3668	0.2637	1.000

Table 3: JASCO vs MusicGen-Melody (Best Configurations)

Model	Configuration	Gen↔Target	Text↔Gen	Melody Acc	Aesthetics CE
JASCO	text_focused (3.0/3.0)	0.4292	0.3035	0.3346	0.978
MusicGen-Melody	guidance_medium (3.0)	0.2146	0.2760	0.2674	1.000

6.3 Model Comparison

7 Analysis

CFG Trade-off: Equal weighting (3.0/3.0) achieves the best balance, heavy melody guidance (7.0/0.5) degrades target matching while heavy text guidance (2.0/4.0) maintains reasonable performance.

JASCO vs MusicGen-Melody: JASCO outperforms MusicGen-Melody by 100% on Gen↔Target (0.43 vs 0.21), demonstrating that separate CFG controls for melody/chords/drums provide superior controllability over single guidance scale.

LP-MusicCaps vs Qwen2-Audio: Qwen2-Audio generates natural, controllable descriptions focusing on melodic character and rhythm, while LP-MusicCaps produces segment-based technical descriptions with repetitive patterns unsuitable for modern text-to-music models.