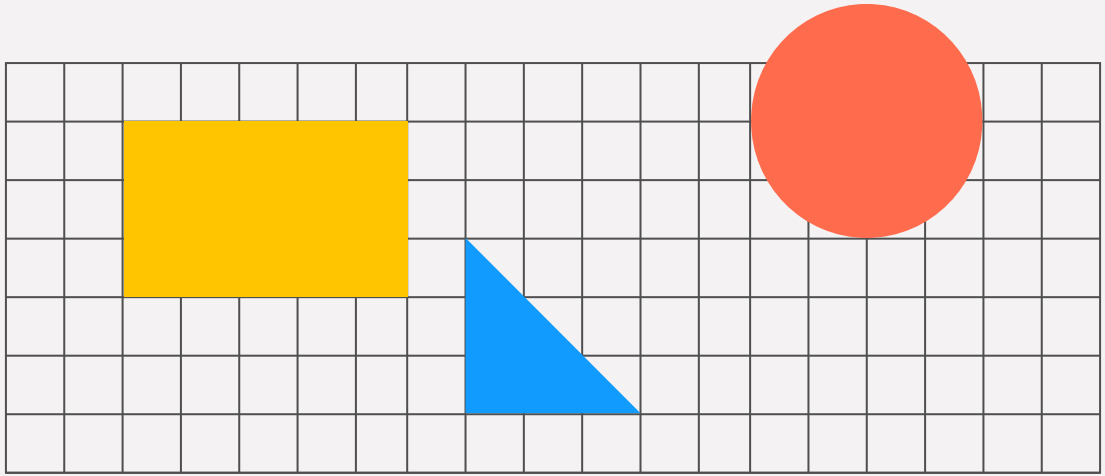
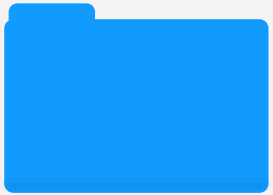
# AlbumGen

Jethro (Cheuk Sau) Au
Joseph (Chun Yu) Lai

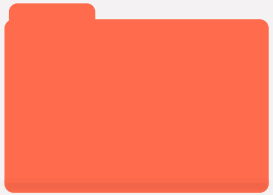


https://github.com/JosephLaiCY/6000N-Multi-modality-LLM

Introduction



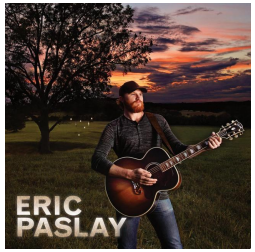
Dataset



Pipeline &
Architecture

Demo &
Conclusion

# Judging Music by Its Cover




**Visual Clues Reflect Musical Themes**

**Colors**: Mood and tone (e.g., dark for metal, vibrant for pop).

**Textures & Styles**: Artistic representation of genre and theme.

**Objects**: Instruments and symbols provide context for the music.

**Project Focus**

Exploring how **LLMs can interpret album art** to generate music.

Connecting **visual elements** to **auditory creativity** using AI.

**Prompt**

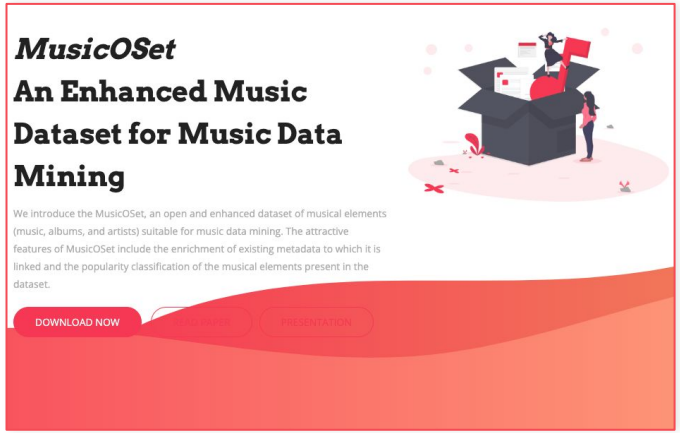
How would songs from this album sound?

Country music?
Mood?

# The Dataset

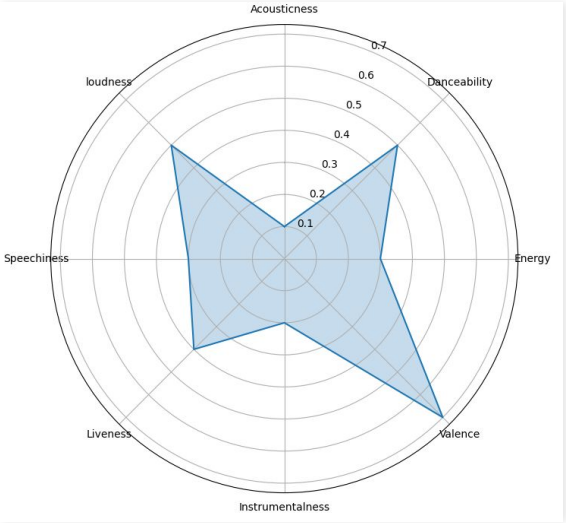
## MusicOSet

MusicOset offers a **comprehensive collection** of enriched metadata related to music, artists, and albums from the U.S. popular music industry. This dataset **is designed for various music data mining tasks**, including visualization, classification, clustering, and similarity searches.

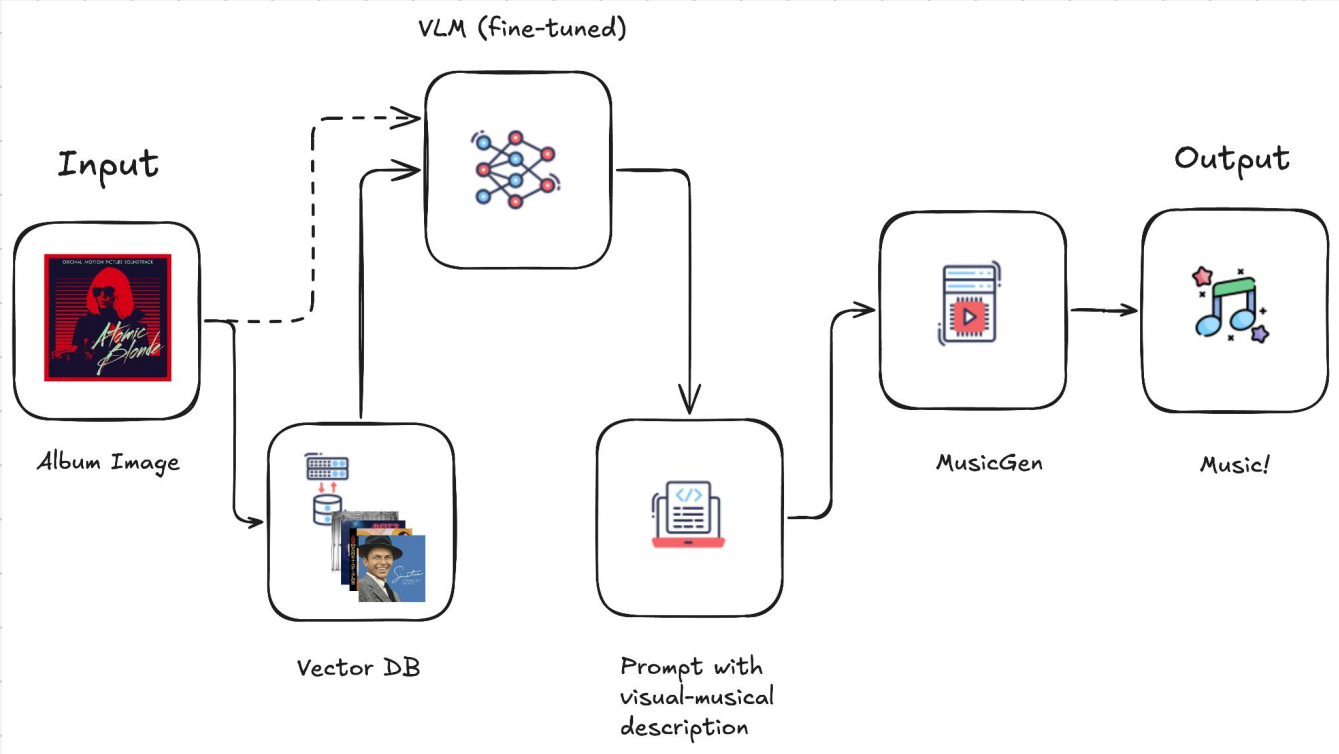


# Musical Attributes

1. Acousticness
2. Danceability
3. Energy
4. Instrumentalness
5. Liveness
6. Loudness
7. Speechiness
8. Valence

# Pipeline & Architecture



# Information Retrieval

**CLIP Model Selection**
Model: **openai/clip-vit-base-patch16** (ViT architecture).
Pre-trained, no fine-tuning, optimized for zero-shot tasks.

**Image Processing**
Dataset: MusicOSet (1,000 album art images).
Resized to 640x640 pixels for consistency.

**Embedding Extraction**
Images passed through CLIP → **512-dimensional embeddings**.
Embeddings capture visual semantics for retrieval and comparison.

# Information Retrieval

## Similarity Metrics

Metric: **L2 Distance** for comparing embeddings (quality over quantity).

## Top K Selection

Dataset split: Development (80%) , validation (10%), test (10%).

Features: **Acousticness**, **danceability**, **energy**, etc.

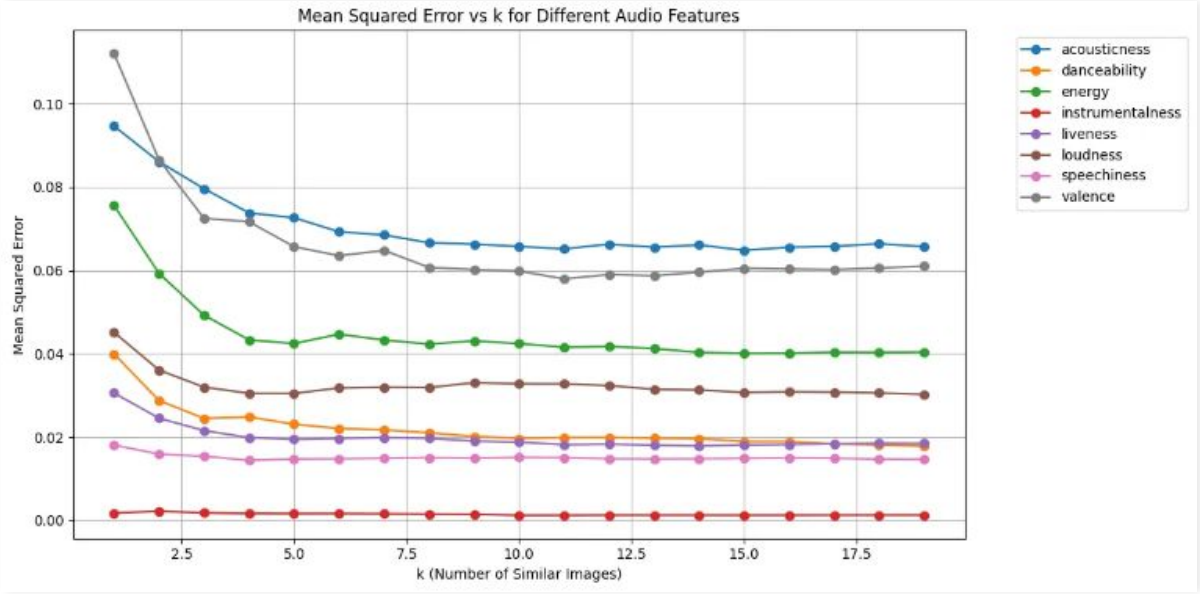
Top K albums: Compare average values of features to ground truth.

Optimal K = **5**: Balances quality and computational efficiency.

## Prompt Creation for MusicGen

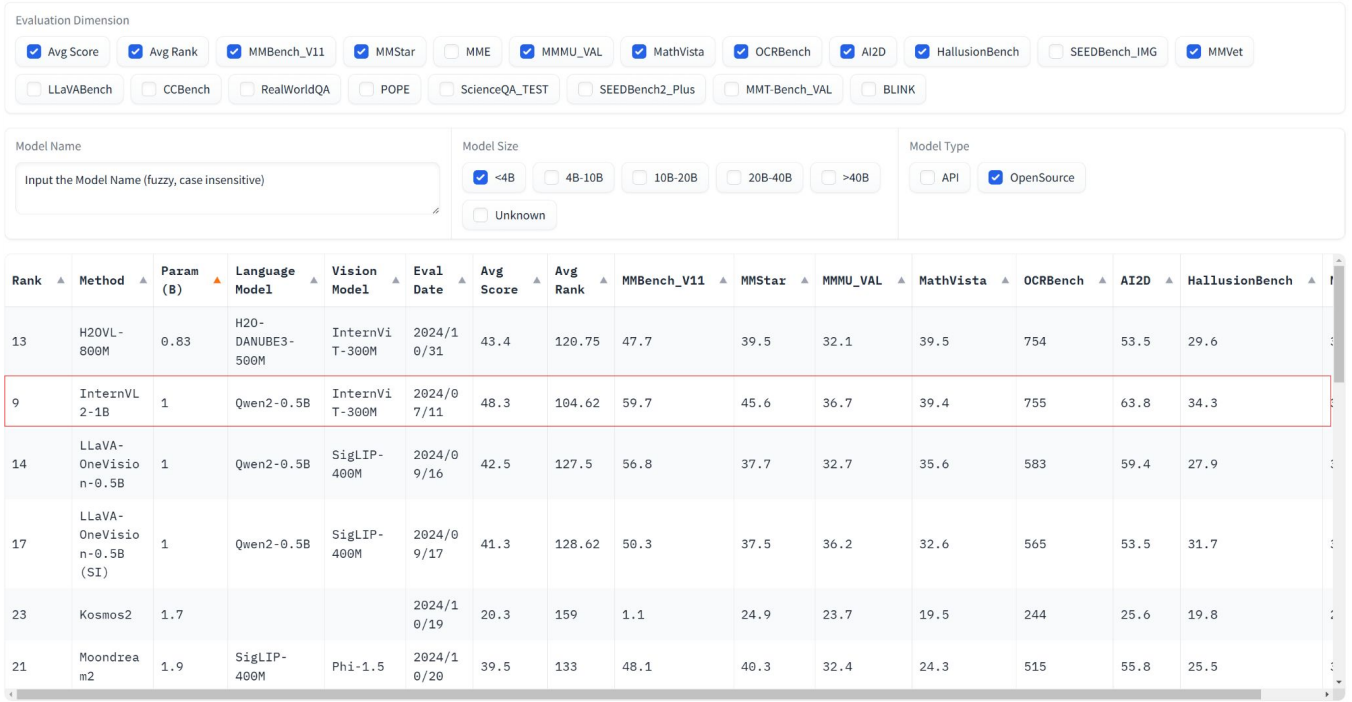
Use average attributes from Top K matches to craft prompts for generating music.

# Evaluation on Retrieval



# Visual LM Selection

## OpenVLM Leaderboard




| Rank | Method | Param (B) | Language Model | Vision Model | Eval Date | Avg Score | Avg Rank | MMBench_V11 | MMStar | MMMU_VAL | MathVista | OCRBench | AI2D | HallusionBench |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | H2OVL-800M | 0.83 | H2O-DANUBE3-500M | InternViT-300M | 2024/10/31 | 43.4 | 120.75 | 47.7 | 39.5 | 32.1 | 39.5 | 754 | 53.5 | 29.6 |
| 9 | InternVL2-1B | 1 | Qwen2-0.5B | InternViT-300M | 2024/07/11 | 48.3 | 104.62 | 59.7 | 45.6 | 36.7 | 39.4 | 755 | 63.8 | 34.3 |
| 14 | LLaVA-OneVision-0.5B | 1 | Qwen2-0.5B | SigLIP-400M | 2024/09/16 | 42.5 | 127.5 | 56.8 | 37.7 | 32.7 | 35.6 | 583 | 59.4 | 27.9 |
| 17 | LLaVA-OneVision-0.5B (SI) | 1 | Qwen2-0.5B | SigLIP-400M | 2024/09/17 | 41.3 | 128.62 | 50.3 | 37.5 | 36.2 | 32.6 | 565 | 53.5 | 31.7 |
| 23 | Kosmos2 | 1.7 | | | 2024/10/19 | 20.3 | 159 | 1.1 | 24.9 | 23.7 | 19.5 | 244 | 25.6 | 19.8 |
| 21 | Moondream2 | 1.9 | SigLIP-400M | Phi-1.5 | 2024/10/20 | 39.5 | 133 | 48.1 | 40.3 | 32.4 | 24.3 | 515 | 55.8 | 25.5 |

# Visual LM:

**Pretrained Model:**

**Model**: OpenGVLab/InternVL2-1B
**Vision Projector:** InternViT-300M-448px
**LLM:** Qwen2-0.5B-Instruct

**Environment:**
Nvidia NGC Pytorch 24.05 Container
RTX3070Ti

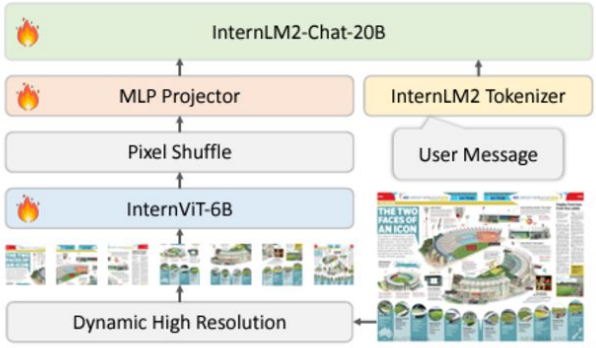



Figure 3. **Overall Architecture.** InternVL 1.5 adopts the ViT-MLP-LLM architecture similar to popular MLLMs [62, 64], combining a pre-trained InternViT-6B [18] with InternLM2-20B [11] through a MLP projector. Here, we employ a simple pixel shuffle to reduce the number of visual tokens to one-quarter.

# VLM Finetuning with Lora

**LoRA Fine Tuning Method:**

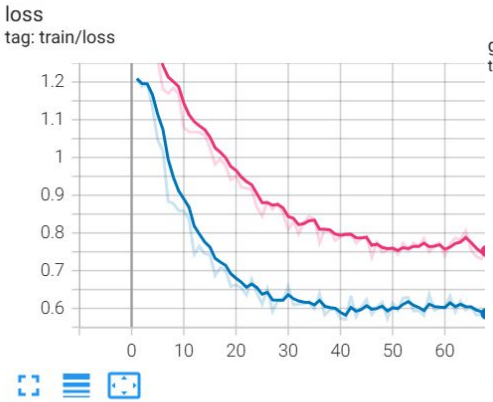
Num examples = 1,102

Number of training steps = 68

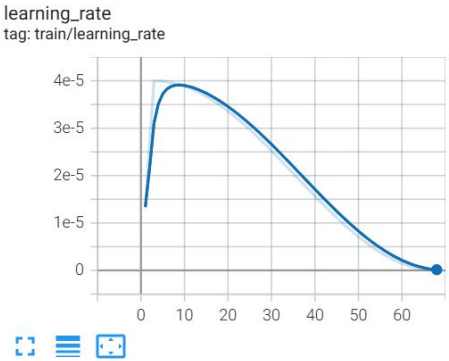
Num Epochs = 1

Total train batch size = 16

Number of trainable parameters = 2,199,552







# MusicGen

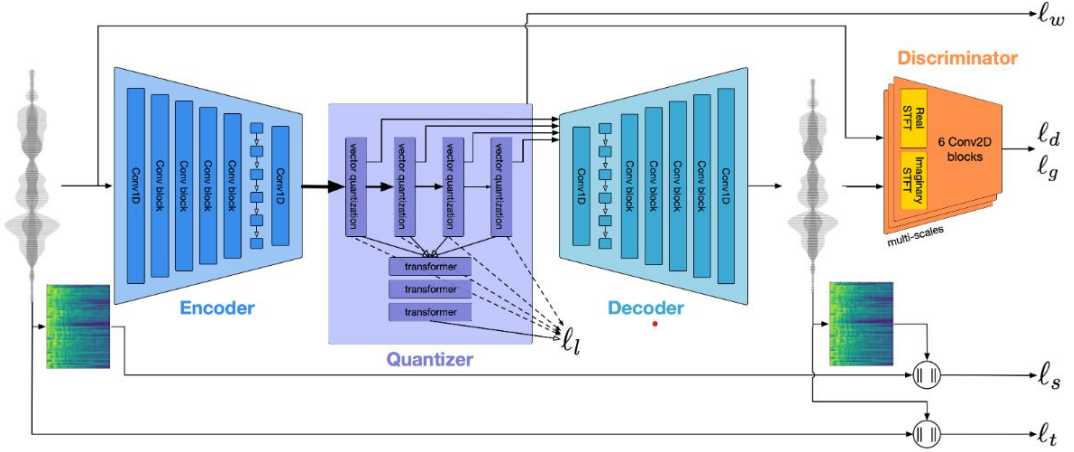
## Simple and Controllable Music Generation

**Jade Copet**[♠◇] **Felix Kreuk**[♠◇] **Itai Gat** **Tal Remez** **David Kant**
**Gabriel Synnaeve** [◇] **Yossi Adi**[◇] **Alexandre Défossez** [◇]
♠: equal contributions, ◇: core team
Meta AI
{jadecopet, felixkreuk, adiyoss}@meta.com

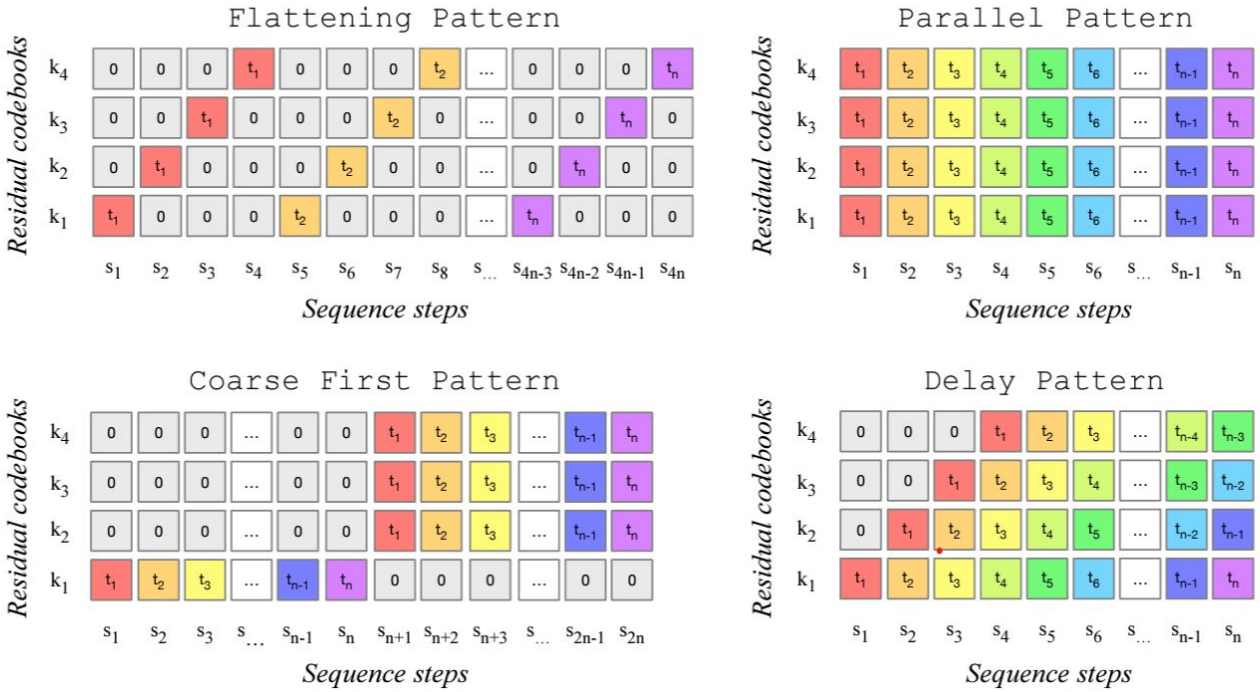
### Abstract

We tackle the task of conditional music generation. We introduce MUSICGEN, a single Language Model (LM) that operates over several streams of compressed discrete music representation, i.e., tokens. Unlike prior work, MUSICGEN is comprised of a single-stage transformer LM together with efficient token interleaving patterns, which eliminates the need for cascading several models, e.g., hierarchically or upsampling. Following this approach, we demonstrate how MUSICGEN can generate high-quality samples, both mono and stereo, while being conditioned on textual description or melodic features, allowing better controls over the generated output. We conduct extensive empirical evaluation, considering both automatic and human studies, showing the proposed approach is superior to the evaluated baselines on a standard text-to-music benchmark. Through ablation studies, we shed light over the importance of each of the components comprising MUSICGEN. Music samples, code, and models are available at github.com/facebookresearch/audiocraft.

**ENCODEC MODEL**



# MusicGen


Flattening Pattern
Residual codebooks
Sequence steps
Parallel Pattern
Residual codebooks
Sequence steps
Coarse First Pattern
Residual codebooks
Sequence steps
Delay Pattern
Residual codebooks
Sequence steps

# Generate Prompt with Acoustic Description

```
# Construct the prompt
music_prompt = (
    f"Compose a {descriptions['album_genre']} song with a tempo of {descriptions['tempo']} similar to {artist_name}'s style. "
    f"in the key of {descriptions['key']} and a {descriptions['time_signature']} time signature. "
    f"The song should have a {descriptions['valence']}, featuring {descriptions['danceability']} rhythms and "
    f"{descriptions['energy']} levels. Create a {descriptions['acousticness']} sound, "
    f"and ensure it is a {descriptions['liveness']} with a {descriptions['loudness']} overall sound level. "
    f"The song should contain {descriptions['speechiness']}."
)
```

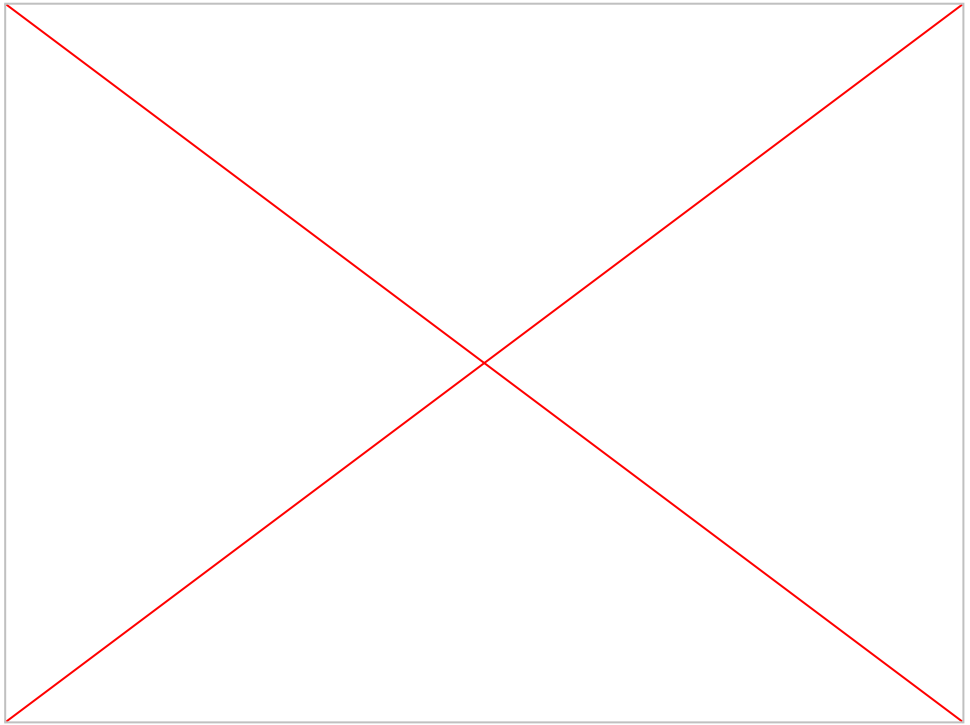
```
descriptions = {
'acousticness': tm.describe_acousticness(acousticness),
'danceability': tm.describe_danceability(danceability),
'energy': tm.describe_energy(energy),
'instrumentalness': tm.describe_instrumentalness(instrumentalness),
'liveness': tm.describe_liveness(liveness),
'loudness': tm.describe_loudness(loudness),
'speechiness': tm.describe_speechiness(speechiness),
'valence': tm.describe_valence(valence),
'tempo': tm.describe_tempo(tempo),
'key': key,
'time_signature': time_signature,
'album_genre': album_genre
}
```

```
def describe_tempo(tempo):
    if tempo > 180:
        return f'very fast tempo ({round(tempo)} BPM)'
    elif tempo > 160:
        return f'fast tempo ({round(tempo)} BPM)'
    elif tempo > 140:
        return f'upbeat tempo ({round(tempo)} BPM)'
    elif tempo > 120:
        return f'moderately fast tempo ({round(tempo)} BPM)'
    elif tempo > 100:
        return f'moderate tempo ({round(tempo)} BPM)'
    elif tempo > 80:
        return f'slow tempo ({round(tempo)} BPM)'
    elif tempo > 60:
        return f'very slow tempo ({round(tempo)} BPM)'
    else:
        return f'extremely slow tempo ({round(tempo)} BPM)'
```

# Demo

**Generated Music Prompt**

Compose a pop rock song with a tempo of very slow tempo (77 BPM) similar to Queen's style. in the key of D and a 30 time signature. The song should have a a very upbeat and cheerful song with a bright and optimistic mood that lifts the spirits, potentially using energetic effects or harmonic layering, featuring a somewhat danceable piece with subtle rhythmic elements, possibly using ambient effects or soft percussion to inspire gentle swaying rhythms and a moderately energetic track balancing intensity with calmer moments, possibly incorporating effects like rhythmic delay or subtle distortion to enhance certain sections levels. Create a a balanced blend of acoustic and electronic sounds, merging natural instruments with modern production techniques and effects like delay or ambient reverb to create a fusion of textures sound, and ensure it is a a clearly produced studio recording, with minimal ambient noise or indaication of a live environment with a an extremely loud track, with intense sound levels and powerful dynamics that command attention overall sound level. The song should contain an almost entirely musical piece, with little to no spoken words, focusing on melodies and harmonies.



# Demo: Output sometimes can be similar


W/out RAG

w/ RAG

# Rock Album Demo: Actual


# Demo: W/OUT RAG


## VLM Output

{"Artist Name": "TOKYO POLICE CLUB",
"Album Genre": "rock",
"key": "C",
"time_signature": "4",
"acousticness": 0.073,
"danceability": 0.7440000000000001,
"energy": 0.958,
"instrumentalness": 0.0,
"liveness": 0.117,
"loudness": 0.334 ,
"speechiness": 0.0152,
"valence": 0.890 ,
"tempo": 116.061}

## Music Prompt

Compose a **rock** song with a tempo of moderate tempo (**116 BPM**) similar to **TOKYO POLICE CLUB**'s style. in the **key of C** and a **4 time signature**.
The song should have a a very upbeat and cheerful song with a bright and optimistic mood that lifts the spirits, potentially using energetic effects or harmonic layering, featuring a quite danceable tune with rhythms and beats that encourage movement, perhaps incorporating electronic elements like synth basslines or rhythmic gating effects rhythms and an extremely energetic and intense track, brimming with high-tempo rhythms and vigorous performances…

# Demo: RAG Example


## VLM Output

{"Artist Name" : "TOKYO POLICE CLUB Forcefield",
"Album Genre": "Disco-Rock",
"key": "C",
"time_signature": "4",
"acousticness": 0.16,
"danceability": 0.52,
"energy": 0.82,
"instrumentalness": 0.000001,
"liveness": 0.13688,
"loudness": 0.721,
"speechiness": 0.10431,
"valence": 0.468,
"tempo": 160.00000000000002}

## Music Prompt

Compose a **Disco-Rock** song with a tempo of fast tempo **(160 BPM)** similar to **TOKYO POLICE CLUB Forcefield**'s style. in the **key of C and a 4 time signature.**
The song should have a a moderately positive piece with a slightly happy mood that is enjoyable and lighthearted, where subtle effects support the pleasant atmosphere without dominating it, featuring a moderately danceable track with a noticeable rhythmic feel, where subtle production effects like reverb on drums contribute to its groove rhythms and a very energetic song, lively and dynamic…

# Rap Demo: Similar prompt but with different result


W/out RAG

w/ RAG

# Demo: W/OUT RAG


## VLM Output


```
{"Artist Name": "RITZ",
"Album Genre": "Hip Hip Hop",
"key": "F",
"time_signature": "4",
"acousticness": 0.00445,
"danceability": 0.8046,
"energy": 0.863,
"instrumentalness": 0.00004,
"liveness": 0,
"loudness": 1.017,
"speechiness": 0.008,
"valence": 0.5427,
"tempo": 139.06}
```


## Music Prompt

Compose a **Hip Hip Hop** song with a tempo of moderately fast tempo **(139 BPM)** similar to **RITZ's** style. in the **key of F and a 4 time signature.** The song should have a a moderately positive piece with a slightly happy mood that is enjoyable and lighthearted, where subtle effects support the pleasant atmosphere without dominating it, featuring a highly danceable song featuring compelling grooves and rhythms, possibly utilizing effects like phasers or flangers on percussion to add movement and energy rhythms and a very energetic song, lively and dynamic, likely featuring powerful vocals and …

# Demo: RAG Example


## VLM Output


```
{"Artist Name": "RIiTZ",
"Album Genre": "r&b",
"key": "A",
"time_signature": "4",
"acousticness": 0.4216,
"danceability": 0.7459,
"energy": 0.8404,
"instrumentalness": 0.0658,
"liveness": 0.160000000000000002,
"loudness": 0.8222,
"speechiness":
0.045000000000000004,
"valence": 0.5877,
"tempo": 76.22}
```


## Music Prompt

Compose a **r&b** song with a tempo of very slow tempo (**76 BPM**) similar to **RIiTZ's style**. in the key of **A and a 4 time signature.**

The song should have a a moderately positive piece with a slightly happy mood that is enjoyable and lighthearted, where subtle effects support the pleasant atmosphere without dominating it, featuring a quite danceable tune with rhythms and beats that encourage movement, perhaps incorporating electronic elements like synth basslines or rhythmic gating effects rhythms and a very energetic song, lively and dynamic, …

# Rap Demo: Actual


# Demo: Very Different Generations


W/out RAG

w/ RAG

# Demo: W/OUT RAG


## VLM Output

{"Artist Name": "Bo Bice",
"Album Genre": "R&B, Country",
"key": "C#",
"time_signature": "4",
"acousticness": 0.2398,
"danceability": 0.2877,
"energy": 0.644,
"instrumentalness": 0.0,
"liveness": 0.63,
"loudness": 0.817,
"speechiness": 0.0,
"valence": 0.895,
"tempo": 152.61}

## Music Prompt

Compose a **R&B, Country song** with a **tempo of upbeat** tempo (**153 BPM**) similar to **Bo Bice's style.** in the **key of C# and a 4 time signature.**
The song should have a a very upbeat and cheerful song with a bright and optimistic mood that lifts the spirits, potentially using energetic effects or harmonic layering, featuring less danceable, featuring rhythmically complex or subdued beats, where experimental effects might create an introspective atmosphere not conducive to dancing rhythms and a quite energetic piece with a strong sense of movement, perhaps using modulation effects like tremolo or vibrato to add excitement levels…

# Demo: RAG (Hallucination)


## VLM Output

{"Artist Name": "BO BICE 3",
"Album Genre": "male",
"key": "C",
"time_signature": "4",
"acousticness":
0.09600000000000001,
"danceability":
0.62200000000000018,
"energy": 0.63400000000000006,
"instrumentalness":
0.058300000000000005,
"liveness": 0.592,
"loudness": 0.70214999999999995,
"speechiness": 0.0621,
"valence": 0.4003,
"tempo": 160.00000000000006}

## Music Prompt

Compose a **male** song with a tempo of fast tempo (**160 BPM**) similar to **BO BICE 3**'s style. in the key of **C and a 4 time signature.**
The song should have a a neutral mood, neither particularly happy nor sad, effects are likely used sparingly and subtly, featuring a quite danceable tune with rhythms and beats that encourage movement, perhaps incorporating electronic elements like synth basslines or rhythmic gating effects rhythms and a quite energetic piece with a strong sense of movement, perhaps using modulation effects like tremolo or vibrato to add excitement levels. Create a a non-acoustic piece …

# Demo: Actual


# Conclusion & Limitations

**Conclusion**

- Textual intermediaries allow for more explicable control of generation between difference modalities
  - Combining Image → text and text → audio
- Finetuning improved  the stability of the generation but can pretrained could still generate just as accurate audio

**Limitations:**

- Variability and accuracy of the music is still limited by the textual descriptions the music can generation
- Furtherwork, directly training the embedding layers amongst music albums & images → if can work around music licensing constraints on data

# Thank you