25-2 DSL Modeling Multi-Modal Wave

13기 한연주, 박시현, 서승범

14기 이재원

1. Introduction

멀티모달 생성 모델은 서로 다른 양식(modality)의 정보를 변환하고 결합하여 새로운 콘텐츠를 생성하는 기술로, 최근 컴퓨터 비전과 오디오 생성 분야를 중심으로 활발히 연구되고 있다. 특히 이미지로부터 텍스트를 추출하고 이를 기반으로 오디오를 생성하는 image-to-audio 변환은 영상 콘텐츠 제작, 접근성 향상, 창작 도구 개발 등 다양한 응용 가능성을 지닌다. 그러나 이미지와 오디오는 정보의 표현 방식과 구조가 근본적으로 상이하다는 점에서, 두 양식 간 의미적 정렬(semantic alignment)을 달성하는 것은 여전히 어려운 과제로 남아 있다.

단순히 Vision-Language Model(VLM)과 오디오 생성 모델을 순차적으로 연결하는 방식은 실용적인 한계를 가진다. VLM이 생성한 텍스트가 오디오 생성 모델의 입력 형식과 목적에 적합하지 않을 경우, 이미지와 무관한 소리가 생성되거나 과도한 노이즈, 혹은 의도하지 않은 음악적 출력이 발생할 수 있다. 또한 이미지에 존재하지 않는 객체나 상황을 텍스트로 생성하는 hallucination 문제가 발생하면, 후속 오디오 생성 단계에서도 오류가 연쇄적으로 증폭된다. 이러한 문제는 개별 모델의 성능 문제라기보다, 모델 간 연결 과정에서 사용되는 중간 표현(intermediate representation)이 충분히 설계되지 않았기 때문에 발생한다.

본 프로젝트는 이러한 문제의식을 바탕으로, 이미지로부터 음악을 생성하는 과정을 단일 end-to-end 모델로 해결하기보다는 **단계별 역할이 명확히 분리된 파이프라인 설계 문제**로 접근한다. 특히 오디오 생성 단계의 특성을 먼저 분석하고, 그에 적합한 텍스트 입력 형식을 역으로 설계하는 backward design 접근법을 채택하였다. 이를 통해 이미지 이해 중심의 텍스트 생성이 아닌, 오디오 생성에 최적화된 의미 표현을 중간 단계로 사용하고자 한다.

궁극적으로 본 프로젝트는 일상 이미지를 입력으로 받아 해당 장면에서 발생할 법한 소리를 생성하고, 이를 음악적 재료로 확장하는 멀티모달 생성 파이프라인을 구축하는 것을 목표로 한다. 이를 위해 image-to-sound와 sound-to-music의 두 단계를 명확히 구분하고, 각 단계에서 요구되는 모델 선택과 설계 기준을 체계적으로 정립하였다.

2. Overall Pipeline

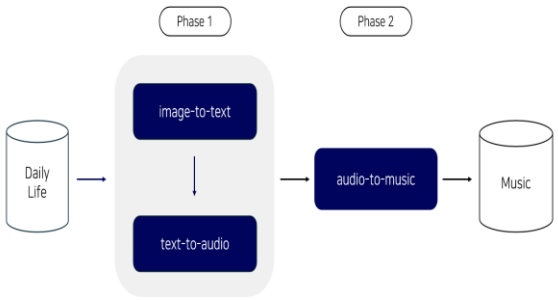

그림 1 전체 파이프라인

본 프로젝트에서 제안하는 전체 파이프라인은 크게 두 개의 단계로 구성된다. 첫 번째 단계는 이미지를 입력으로 받아 해당 장면에 어울리는 효과음을 생성하는 **Phase 1: Image-to-Sound** 단계이며, 두 번째 단계는 생성된 효과음을 악기 음색으로 변환하고 음악으로 재구성하는 **Phase 2: Sound-to-Music** 단계이다.

Phase 1에서는 이미지에 대한 장면(scene)과 분위기(mood)를 분석하고, 이미지 속 객체와 그 재질 및 상호작용을 기반으로 소리를 낼 수 있는 sound source를 텍스트 형태로 설계한다. 이후 해당 텍스트를 입력으로 하여 text-to-audio 모델을 통해 현실적인 효과음을 생성한다. 이 단계의 출력은 음악이 아닌 비음악적(sound effect) 오디오로, Phase 2에서 음악적 변환을 수행하기 위한 입력으로 사용된다.

Phase 2에서는 Phase 1에서 생성된 효과음을 음악적 재료로 활용한다. 구체적으로는 효과음의 음색을 특정 악기 음색으로 변환하고, 이를 음계 전반에 매핑하여 악기별 soundfont를 생성한다. 이후 MIDI 악보를 기반으로 해당 soundfont를 연주함으로써 최종 음악을 생성한다. 이 과정은 작곡이나 편곡을 자동으로 생성하는 것이 아니라, 이미지에서 유래한 소리를 악기 음색으로 재해석하여 기존 음악 구조에 반영하는 데 초점을 둔다.

이와 같은 구조에서 본 프로젝트는 이미지에서 음악으로의 변환을 단일 모델의 출력으로 처리하지 않는다. 대신, 각 단계에서 해결해야 할 문제를 분리하고, 단계 간 전달되는 중간 표현을 명시적으로 설계함으로써 의미적 일관성과 제어 가능성을 확보하고자 한다. 이러한 모듈형 파이프라인 설계는 개

별 단계의 개선이나 대체가 용이하다는 장점도 함께 제공한다.

3. Phase 1: Image-to-Sound

3.1 Phase 1 Overview

Phase 1의 목적은 입력 이미지에 어울리는 **현실적인 효과음(sound effect)**을 생성하는 것이다. 본 단계에서 생성되는 오디오는 음악이 아닌, 이미지 속 장면에서 실제로 발생할 법한 비음악적 소리로 정의된다. 이러한 효과음은 이후 Phase 2에서 음악적 변환을 수행하기 위한 입력으로 활용된다.

Phase 1은 이미지를 직접 오디오로 변환하는 단순한 매핑 문제가 아니라, 이미지로부터 소리를 설계하기 위한 의미적 중간 표현을 구성하는 과정으로 이해할 수 있다. 이를 위해 본 프로젝트는 이미지에 대한 장면(scene)과 분위기(mood)를 분석하고, 이미지 속 객체와 그 재질, 상호작용을 기반으로 소리를 낼 수 있는 sound source를 텍스트 형태로 구조화한다. 이후 해당 텍스트를 입력으로 하여 text-to-audio 생성 모델을 통해 효과음을 생성한다.

요약하면, Phase 1은 image-to-text와 text-to-audio의 두 단계로 구성되며, 핵심 목표는 이미지와 오디오 간의 의미적 일관성을 유지하면서도 Phase 2에서 활용 가능한 형태의 효과음을 안정적으로 생성하는 데 있다.

3.2 Design Motivation & Approach

이미지로부터 효과음을 생성하기 위해 Vision-Language Model(VLM)과 text-to-audio 모델을 순차적으로 연결하는 접근은 직관적으로 보이지만, 실제로는 다양한 문제를 야기한다. VLM이 생성한 텍스트가 오디오 생성 모델의 입력 형식이나 목적에 부합하지 않을 경우, 이미지와 무관한 소리가 생성되거나 과도한 노이즈, 혹은 음악적 요소가 포함된 출력이 발생할 수 있다. 또한 VLM이 이미지에 존재하지 않는 객체를 텍스트로 생성하는 hallucination 문제가 발생하면, 이후 오디오 생성 단계에서도 오류가 증폭된다.

이러한 문제를 해결하기 위해 본 프로젝트는 **backward design 접근법**을 채택하였다. 즉, 이미지 이해에서 출발하는 것이 아니라, 먼저 text-to-audio 모델이 어떤 형식과 내용의 텍스트 입력에서 가장 안정적이고 현실적인 효과음을 생성하는지를 분석하고, 이를 기준으로 VLM의 출력 형식을 설계하였다. 다시 말해, Phase 1에서는 image-to-text 모델이 생성해야 할 텍스트를 사전에 정의하고, 그 요구사항에 맞춰 전체 파이프라인을 구성하였다.

이 접근법을 통해 VLM의 출력은 단순한 이미지 설명이나 캡션이 아니라, 오디오 생성에 직접 활용 가능한 **audio-oriented representation**으로 기능하게 된다. 이는 Phase 1의 성능을 개별 모델의 성능이 아닌, 파이프라인 설계 문제로 다루고자 한 본 프로젝트의 핵심적인 설계 철학이다.

3.3 Dataset (Image-to-Sound)

Phase 1에서 사용된 데이터셋은 VLM의 이미지 이해 능력을 다각도로 평가하기 위해 선정되었다. 본 프로젝트에서는 일상적인 객체 인식 능력과 추상적인 분위기 인식 능력을 모두 고려하여 두 가지 공개 데이터셋을 활용하였다.

첫 번째 데이터셋은 MSCOCO(Microsoft Common Objects in Context)로, 다양한 일상 장면과 명확한 객체 정보를 포함한 이미지들로 구성되어 있다. MSCOCO는 VLM이 이미지 속 객체를 정확히 식별하고, 해당 객체의 재질이나 물리적 특성을 추론하는 능력을 평가하는 데 적합하다.

두 번째 데이터셋은 ArtEmis로, 예술 회화 이미지를 중심으로 감정적 반응과 분위기 정보를 포함한다. ArtEmis는 명확한 객체가 존재하지 않거나 추상적인 장면에서도 이미지의 전반적인 무드와 정서를 파악하는 능력을 평가하기 위해 사용되었다.

각 데이터셋에서 무작위로 11장의 이미지를 선택하여 총 22장의 이미지로 평가용 데이터셋을 구성하였다. 본 프로젝트의 목적이 모델 학습이 아닌 프롬프트 설계와 모델 평가에 있기 때문에, 데이터셋의 규모는 제한적으로 설정되었다. 대신, 각 이미지에 대해 사람의 직관적 인식을 기반으로 한 Golden Dataset을 구축하여 평가 기준으로 활용하였다.

Golden Dataset은 이미지별로 장면 설명(scene description), 분위기 설명(mood description), 그리고 소리를 생성할 수 있는 객체와 그 속성을 구조화한 sound source 정보를 포함한다. 이 데이터는

이후 VLM 출력의 정확성과 일관성을 평가하는 기준으로 사용되었다.

3.4 Text-to-Audio Model Selection

Phase 1에서 효과음을 생성하기 위한 text-to-audio 모델을 선정하기 위해, 여러 후보 모델을 대상으로 비교 실험을 수행하였다. 대표적으로 Make-an-Audio 계열 모델과 AudioLDM2를 동일한 프롬프트 조건에서 비교하였다.

실험 결과, Make-an-Audio 모델은 이미지와 직접적으로 연관되지 않은 소리를 생성하거나, 노이즈가 심해 생성된 오디오의 구분이 어려운 경우가 빈번하게 관찰되었다. 반면 AudioLDM2는 프롬프트에 명시된 장면과 분위기를 비교적 잘 유지하며, 전반적으로 노이즈가 적고 현실적인 앰비언스에 가까운 효과음을 생성하는 경향을 보였다.

또한 Phase 1의 목적이 음악 생성이 아닌 효과음 생성이라는 점을 고려할 때, AudioLDM2는 프롬프트 제약을 통해 비음악적 출력을 보다 안정적으로 유도할 수 있다는 장점을 지녔다. 이러한 정성적 비교 결과를 바탕으로, Phase 1의 text-to-audio 모델로 AudioLDM2를 최종 선정하였다.

3.5 Prompt Design for AudioLDM2

AudioLDM2를 효과음 생성에 적합하게 활용하기 위해, 본 프로젝트에서는 프롬프트 구조를 체계적으로 설계하였다. 프롬프트는 크게 공통 프롬프트와 세부 프롬프트로 구성된다. 공통 프롬프트에는 이미지의 장면과 분위기에 대한 설명을 포함하여, 생성될 오디오의 전반적인 맥락을 제공한다.

세부 프롬프트에서는 소리를 생성할 객체, 해당 객체의 재질, 소리 발생 방식(play method), 음색(timbre) 정보를 명시한다. 또한 특정 sound source가 이후 Phase 2에서 어떤 악기 세션과 연결될 수 있는지를 나타내는 매핑 정보를 선택적으로 포함하였다. 다만, 매핑 정보가 과도하게 오디오의 질을 저해하지 않도록, 매핑이 불확실한 경우에는 None으로 처리하였다.

프롬프트 구성 요소의 중요성을 검증하기 위해 ablation study를 수행한 결과, 장면 및 분위기 정보는 생성된 소리의 톤과 일관성에 큰 영향을 미쳤으며, 비음악적 효과음을 생성하라는 제약이 없는 경우 모델이 음악적 출력을 생성하는 경향이 확인되었다. 이를 통해 프롬프트 제약 조건이 Phase 1의 성능에 핵심적인 역할을 한다는 점을 확인하였다.

3.6 Image-to-Text Model (VLM) Selection

Phase 1에서 VLM은 이미지로부터 AudioLDM2에 적합한 텍스트 출력을 생성하는 역할을 수행한다. 이를 위해 본 프로젝트에서는 LLaVA-NEXT-7B와 Qwen2.5-VL-7B 두 모델을 후보로 선정하여 비교 평가를 진행하였다.

평가는 앞서 구축한 Golden Dataset을 기준으로 수행되었으며, 각 이미지에 대해 생성된 장면 설명, 분위기 설명, sound source 정보가 사람의 직관적 인식과 얼마나 일치하는지를 중심으로 평가하였다. 특히 이미지에 존재하지 않는 객체를 생성하는 hallucination 여부를 중요한 평가 기준으로 삼았다.

평가 결과, LLaVA는 장면 설명 단계에서 hallucination이 빈번하게 발생하였으며, 이는 이후 sound source 설계에서도 부정적인 영향을 미쳤다. 반면 Qwen2.5-VL-7B는 전반적으로 hallucination이 적고, 생성된 텍스트가 AudioLDM2의 입력으로 보다 안정적으로 활용 가능한 형태를 유지하였다. 이에 따라 Phase 1의 image-to-text 모델로 Qwen2.5-VL-7B를 최종 선정하였다.

3.7 Phase 1 Output

Phase 1의 최종 출력은 구조화된 sound source 텍스트와 이를 기반으로 생성된 효과음 오디오 파일이다. sound source는 JSON 형태로 저장되어 장면, 분위기, 객체 정보, 음색 및 매핑 정보를 포함하며, 생성된 오디오는 wav 파일 형태로 저장된다.

이러한 출력은 Phase 2에서 음색 변환과 음악 생성을 수행하기 위한 입력으로 사용된다. 즉, Phase 1은 이미지에서 직접 음악을 생성하는 단계가 아니라, 이미지로부터 음악적 변환이 가능한 음향 재료를 설계하고 생성하는 역할을 수행한다.

4. Phase 2: Sound-to-Music

4.1 Phase 2 Overview

Phase 2의 목적은 Phase 1에서 생성된 효과음을 음

악적 구성 요소로 변환하여 최종 음악을 생성하는 것이다. Phase 1이 이미지로부터 현실적인 소리를 설계하고 생성하는 단계라면, Phase 2는 이러한 소리를 악기 음색으로 재해석하고 음악 구조 안에 통합하는 단계로 볼 수 있다.

본 단계에서는 소리의 의미적 출처보다는 음색(timbre)과 주파수 특성에 초점을 맞춘다. 즉, Phase 2는 새로운 멜로디나 작곡을 생성하는 것이 아니라, 기존 음악 구조를 유지한 채 이미지에서 유래한 소리를 악기 음색으로 변환하여 재생하는 것을 목표로 한다. 이를 통해 이미지–소리–음악 간의 연결이 실제 음악 출력으로 이어질 수 있음을 확인하고자 한다.

4.2 Dataset (Sound-to-Music)

Phase 2에서는 악기 음색 변환을 위해 NSynth 데이터셋을 사용하였다. NSynth는 다양한 악기의 단일 음을 고해상도로 녹음한 데이터셋으로, 음색 학습과 변환에 적합한 구조를 갖는다. 본 프로젝트에서는 NSynth 데이터셋 중 keyboard, bass, electric guitar에 해당하는 악기군을 선택하여 사용하였다.

각 악기군에 대해 데이터셋은 학습(train), 검증(validation), 테스트(test) 세트로 분리되어 있으며, 이를 통해 모델이 특정 악기의 음색 특성을 학습하고 새로운 입력 음원에 대해 해당 악기 음색으로 변환할 수 있도록 구성하였다. 데이터 전처리는 RAVE 모델의 입력 요구사항에 맞추어 샘플링 레이트 및 오디오 길이를 통일하는 방식으로 진행되었다.

4.3 Model

Phase 2의 모델링 과정은 음색 변환, 악기 음원 구성, 음악 렌더링의 세 단계로 이루어진다. 각 단계는 독립적인 역할을 수행하지만, 최종 음악 출력을 위해 순차적으로 연결된다.

4.3.1 RAVE

음색 변환을 위해 본 프로젝트에서는 RAVE(Realtime Audio Variational autoEncoder) 모델을 사용하였다. RAVE는 오디오 신호를 잠재 공간(latent space)으로 인코딩한 후 이를 다시 디코딩하는 구조를 가지며, 실시간 오디오 합성과 음색 변환에 적합한 모델이다.

모델 학습은 Variational Autoencoder(VAE) 기반의 표현 학습 단계와, 오디오 품질을 개선하기 위한 adversarial fine-tuning 단계로 구성된다. 이를 통해 입력 오디오의 시간적 구조를 유지하면서도 목표 악기의 음색 특성을 효과적으로 반영할 수 있다.

본 프로젝트에서는 NSynth 데이터셋을 활용하여 keyboard, bass, electric guitar 각각에 대해 별도의 RAVE 모델을 학습하였다. 이후 Phase 1에서 생성된 효과음을 입력으로 하여, 해당 소리를 각 악기 음색으로 변환하는 inference를 수행하였다.

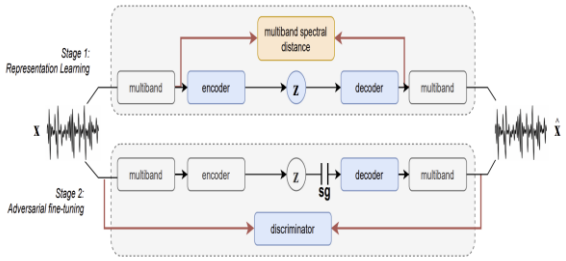

그림 2 RAVE 모델의 훈련 구조

4.3.2 SoundFont Generation

RAVE를 통해 변환된 악기 음색 오디오는 MIDI 기반 음악 렌더링에 활용하기 위해 SoundFont 형식으로 재구성되었다. 이를 위해 librosa 라이브러리를 사용하여 음원의 피치를 조정하고, 단일 음원으로부터 여러 음높이에 대응하는 샘플을 생성하였다.

생성된 샘플들은 autosfz_builder를 통해 자동으로 SoundFont 구조로 정리되었으며, 이를 통해 각 악기별로 음계 전반을 커버하는 가상 악기를 구성할 수 있었다. 드럼 사운드의 경우, 음높이 개념보다는 타격 이벤트 중심의 특성을 고려하여 별도의 피치 변환 없이 키 매핑 방식으로 처리하였다.

이 과정은 환경음이나 효과음과 같은 비전통적 오디오를 악기 음원으로 재해석하는 핵심 단계로, Phase 2에서 음악 생성이 가능하도록 만드는 기반을 제공한다.

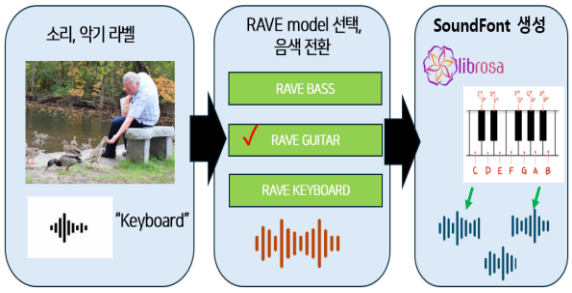

그림 3 SoundFont 생성 과정

4.3.3 MIDI Rendering

최종 음악 생성 단계에서는 PrettyMIDI 라이브러리를 사용하여 MIDI 파일을 렌더링하였다. MIDI 파일은 기존에 구성된 음악 구조를 담고 있으며, 앞서 생성한 SoundFont를 각 트랙에 매핑하여 재생하는 방식으로 음악을 생성하였다.

이 과정에서 멜로디, 베이스, 코드, 드럼 트랙은 각각 해당하는 SoundFont를 사용하여 연주되며, 결과적으로 이미지에서 유래한 소리가 악기 음색으로 반영된 음악이 출력된다. 출력 결과는 단일 wav 파일 형태로 저장되며, 각 악기 트랙이 혼합된 최종 음악을 포함한다.

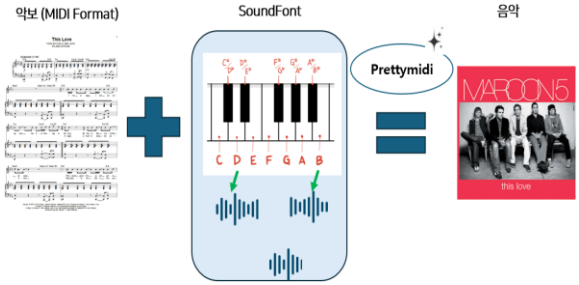

그림 4 SountFont를 이용한 최종 노래 생성 과정

5. Results

5.1 Phase 1 Results: Image-to-Sound

Phase 1의 결과는 생성된 효과음이 입력 이미지의 장면과 의미적으로 얼마나 일관되는지를 중심으로 평가하였다. 정량적 지표보다는, sound source 설계의 타당성과 hallucination 발생 여부, 그리고 생성된 오디오의 현실성을 중심으로 정성적 평가를 수행하였다.

우선 backward design 접근을 통해 설계된 프롬프트 구조는 VLM이 생성한 텍스트를 AudioLDM2의 입력으로 안정적으로 활용할 수 있게 하였다. 장면과 분위기 정보를 명시적으로 포함한 경우, 생성된 효과음은 이미지 전반의 분위기와 잘 부합하는 경향을 보였으며, 객체 단위의 세부 프롬프트를 포함했을 때 소리의 구체성이 향상되었다. 반면 이러한 제약을 제거한 경우에는 음악적 요소가 포함되거나, 장면과 무관한 소리가 생성되는 사례가 관찰되었다.

VLM 비교 실험 결과, Qwen2.5-VL-7B를 사용한 경우 이미지에 존재하지 않는 객체를 생성하는 hallucination 빈도가 상대적으로 낮았으며, 생성된 sound source가 Golden Dataset과 높은 수준의 일관성을 보였다. 이를 통해 Phase 1에서 image-to-text 모델의 선택이 이후 오디오 생성 품질에 직접적인 영향을 미친다는 점을 확인하였다.

종합적으로 Phase 1은 이미지 기반 장면 이해를 바탕으로, Phase 2에서 활용 가능한 형태의 효과음을 안정적으로 생성하는 데 성공하였다.

생성된 결과는 첨부된 링크 (https://github.com/DataScience-Lab-Yonsei/25-2_DSL_Modeling_MultiModal_WAVE/tree/main/sound_to_music/sound_input)에서 확인 가능하다.

5.2 Phase 2 Results: Sound-to-Music

Phase 2에서는 Phase 1에서 생성된 효과음을 입력으로 사용하여, 해당 소리가 음악적 구성 요소로 활용될 수 있는지를 확인하였다. 본 단계의 결과는 생성된 음악의 완성도나 작곡 품질을 평가하기보다는, 효과음이 악기 음색으로 변환되고 MIDI 기반 음악 렌더링 과정에 적용 가능한지를 중심으로 정리한다.

RAVE를 활용한 음색 변환 과정에서는 Phase 1에서 생성된 효과음이 입력으로 사용되었으며, 변환된 오디오는 이후 SoundFont 생성 단계에 활용되었다. 이 과정에서 입력 효과음의 시간적 구조가 유지된 상태로 악기 음색이 적용된 오디오가 생성되었으며, 해당 결과는 SoundFont 구성을 위한 음원으로 사용되었다.

SoundFont 생성 단계에서는 변환된 오디오를 기반으로 여러 음높이에 대응하는 샘플을 구성하여, MIDI 연주에 사용할 수 있는 악기 음원을 생성하였다. 이를 통해 단일 효과음이 음악적 연주에 활용 가능한 형태로 변환될 수 있음을 확인하였다.

생성된 결과는 첨부된 링크 (https://github.com/DataScience-Lab-Yonsei/25-2_DSL_Modeling_MultiModal_WAVE/tree/main/sound_to_music/sfz_output)에서 확인 가능하다.

최종적으로 PrettyMIDI를 사용한 렌더링 과정에서는 생성된 SoundFont가 MIDI 파일에 매핑되어 음악이 출력되었다.

생성된 결과는 첨부된 링크 (https://github.com/DataScience-Lab-Yonsei/25-2_DSL_Modeling_MultiModal_WAVE/tree/main/sound_to_music/result)에서 확인 가능하다.

6. Conclusion

본 프로젝트는 이미지를 입력으로 받아 음악을 생성하는 멀티모달 생성 문제를 단일 모델의 end-to-end 학습 문제가 아닌, **단계별 역할이 분리된 파이프라인 설계 문제**로 접근하였다. 이를 위해 image-to-sound와 sound-to-music의 두 단계를 명확히 구분하고, 각 단계에서 요구되는 중간 표현과 모델 설계 기준을 체계적으로 구성하였다.

Phase 1에서는 이미지로부터 장면과 분위기를 해석하고, 해당 장면에서 발생할 법한 효과음을 생성하는 과정을 다루었다. 이 과정에서 text-to-audio 모델의 입력 요구사항을 기준으로 VLM의 출력 형식을 설계하는 backward design 접근법을 적용함으로써, 이미지와 오디오 간의 의미적 불일치를 완화하고자 하였다. 그 결과, 이미지에 어울리는 효과음을 안정적으로 생성하고, 이후 단계에서 활용 가능한 형태의 오디오 출력을 얻을 수 있었다.

Phase 2에서는 Phase 1에서 생성된 효과음을 음악적 구성 요소로 변환하는 과정을 수행하였다. 음색 변환 모델과 SoundFont 생성, MIDI 렌더링을 통해 효과음이 음악적 연주에 사용될 수 있음을 확인하였으며, 이를 통해 이미지에서 유래한 소리가 음악 생성 과정의 입력으로 활용될 수 있음을 실험적으로 보여주었다.

종합하면, 본 프로젝트는 이미지–소리–음악으로 이어지는 멀티모달 생성 파이프라인을 구현하고, 중간 표현 설계의 중요성과 단계별 접근의 유효성을 확인하였다. 이는 이미지 기반 오디오 및 음악 생성 문제를 다룰 때, 모델 성능뿐만 아니라 파이프라인 설계 자체가 중요한 요소임을 시사한다.

7. Limitations & Future Work

본 프로젝트는 멀티모달 생성 파이프라인의 가능성을 탐색하는 데 목적이 있었으나, 몇 가지 한계점 또한 존재한다.

먼저 Phase 1에서 생성된 sound source와 Phase 2에서 사용되는 악기 세션 간의 매핑은 완전한 자동화가 이루어지지 않았다. 일부 경우에는 동일한 역할의 소리가 중복되거나, 특정 악기 세션이 비어 있는 문제가 발생할 수 있다. 이러한 한계는 향후 sound source의 역할 분류를 보다 정교하게 설계하거나, 세션 수를 유연하게 조정하는 방식으로 개선될 수 있을 것이다.

또한 Phase 2의 SoundFont 생성 과정에서는 제한된 수의 샘플과 단순한 피치 변환 방식을 사용하였다. 이로 인해 음색의 표현력이 제한되며, 실제 악기 연주와 비교했을 때 자연스러움이 부족할 수 있다. 향후에는 보다 다양한 샘플 구성이나 고급 음원 합성 기법을 도입하여 SoundFont의 품질을 개선할 수 있을 것으로 기대된다.

마지막으로 본 프로젝트에서 생성된 음악은 기존 MIDI 구조를 기반으로 렌더링되었으며, 자동 작곡이나 편곡 기능은 포함하지 않았다. 이는 이미지에서 유래한 소리를 음악적으로 어떻게 배치할 것인가에 대한 문제를 단순화하기 위한 선택이었으나, 향후에는 이미지의 의미 정보를 음악 구조나 전개에 반영하는 방향으로 확장할 수 있을 것이다.

이러한 한계에도 불구하고, 본 프로젝트는 멀티모달 생성 문제를 단계적으로 분해하고 각 단계의 역할을 명확히 정의함으로써, 이미지 기반 음악 생성이라는 복합적인 문제를 구조적으로 탐색했다는 점에서 의의를 가진다.