MusicLM

Generating Music From Text

텍스트 설명으로 고품질 음악을 생성

Agostinelli et al., Google Research · ICASSP 2023

발표 목차

연구 배경

텍스트→이미지 생성의 성공이 텍스트→음악 연구를 촉진

텍스트 → 이미지

DALL-E 2, Stable Diffusion

텍스트 설명만으로 고품질 이미지 생성에 성공

→ 자연어 이해 + 고품질 생성의 조합이 다른 도메인에서도 가능하다는 영감

기존 오디오 연구
- TTS: 텍스트 → 음성
- 가사 → 음악

영감 →

텍스트 → 음악

MusicLM (Google, 2023)

텍스트 설명으로 최대 5분의 24kHz 고품질 음악 생성

→ 복잡한 다악기 구조 + 장기적 일관성을 처음으로 달성

핵심 차별점
- 흥얼거림 → 멜로디 유지 변환
- 텍스트 없는 음악 데이터로 학습

기존 모델의 두 가지 핵심 한계

❶ 복잡한 구조 생성의 어려움

단순한 효과음 생성은 가능했지만

여러 악기가 어우러지는 음악
장기적 구조
전반적으로 일관된 흐름

→ 기존 모델로는 일관성 있는
생성이 불가능했습니다

❷ 음악-텍스트 쌍 데이터 부족

지도학습에 필요한 것:
음악 파일 ↔ 텍스트 설명의 대규모 pair 데이터

현실:
텍스트 없이 음악만 있는 데이터는
방대하지만, 텍스트+음악 쌍은 극히 부족

→ **MusicLM은 MuLan으로 우회**
텍스트 없이도 학습 가능!

MusicLM의 핵심

세 가지 혁신적 접근으로 기존 한계를 극복

①

계층적 생성

AudioLM 구조를 확장
Semantic → Coarse Acoustic → Fine Acoustic
3단계 분업으로 24kHz 고음질과 일관성 동시 확보

②

MuLan 조인트 임베딩

텍스트-오디오 공유 공간(128차원)
학습: 28만 시간의 오디오만 사용
추론: 텍스트 입력만으로 음악 생성

③

멜로디 컨디셔닝

흥얼거림·휘파람 오디오 입력을 지원
Semantic Token으로 멜로디 뼈대만 추출
텍스트로 지정한 스타일로 변환

Part 2

핵심 구성요소 3가지

SoundStream · w2v-BERT · MuLan

각각 독립적으로 사전 학습된 신경망(Pretrained & Frozen)

AudioLM — MusicLM의 직접적 조상

MusicLM은 AudioLM의 계층적 구조를 텍스트 조건부 생성으로 확장한 모델

AudioLM
Google, 2022

계층적 구조 계승

MusicLM
+ 텍스트 조건 (MuLan)

Semantic Tokens (S)

모델: w2v-BERT

역할: 음악의 장기적 구조와 멜로디 담당
인트로 → 버스 → 코러스 같은 흐름 유지

특징: 음색·질감은 무시하고
'무슨 음악인가'의 뼈대만 인코딩

Acoustic Tokens (A)

모델: SoundStream (RVQ)

역할: 음색, 질감 등 세부 음향 특성 담당
Semantic 위에 올라가 실제 소리를 완성

특징: 24kHz 고음질 복원 가능
계층적 RVQ로 Coarse/Fine 분리

AudioLM = 텍스트 조건 없이 오디오만으로 생성 | MusicLM = AudioLM + MuLan(텍스트 조건) 추가

① MuLan

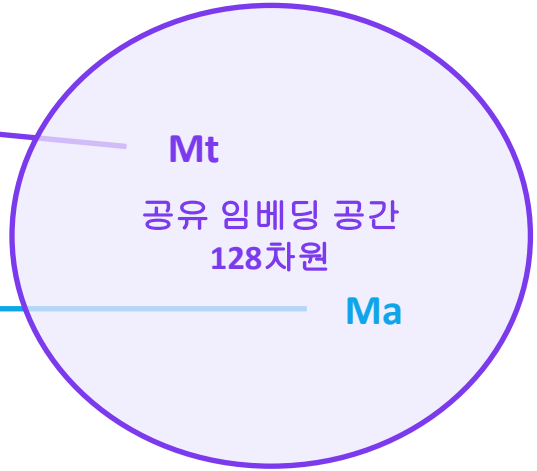
조건부 토큰 (M) — 텍스트와 오디오를 연결하는 공유 임베딩

텍스트 인코더

"재즈 피아노, 빠른 템포"

오디오 인코더

♪ 음악 오디오 파일



🏋 학습 시

텍스트 없이 음악 오디오만으로 학습
Ma를 조건으로 Semantic Token 예측

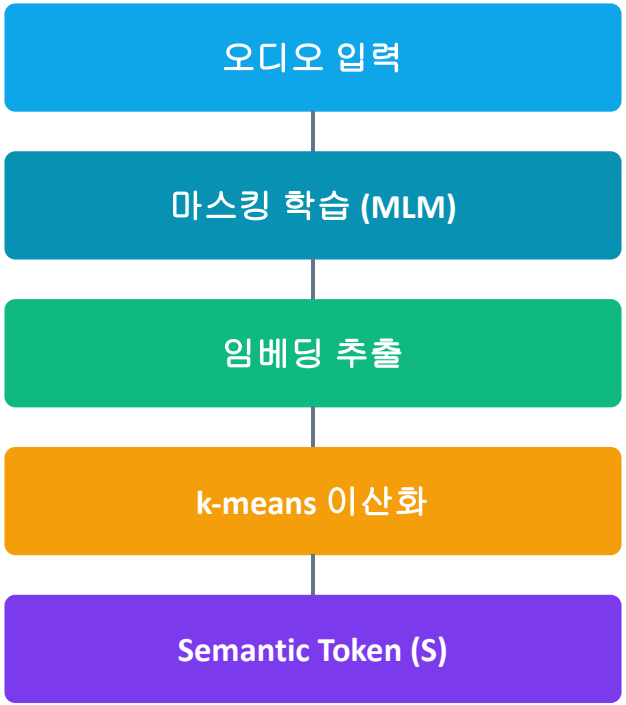
🚀 추론(Inference) 시

사용자 텍스트 → Mt 추출
Ma와 같은 공간에 있어 그대로 사용!

학습은 오디오만으로 · 추론은 텍스트로 → 데이터 부족 문제를 해결

② w2v-BERT

의미 토큰 (Semantic Tokens, S) — 멜로디·리듬·장기 구조 담당



k-means 이산화 원리

학습 단계
방대한 음악 임베딩들을 공간에 뿌린 후 가장 대표적인 1024개 중심점(Centroid) 탐색

이산화 단계
새 임베딩이 들어오면 1024개 중심점 중 가장 가까운 번호를 할당

효과
미세한 노이즈와 불필요한 정보는 버리고 핵심 음악 특징만 남김

600M 파라미터 · 7번째 레이어 임베딩 · k-means 1024 클러스터

③ SoundStream + RVQ

음향 토큰 (Acoustic Tokens, A) — 음색과 질감 담당

SoundStream이란?

오디오 파일을 숫자로 압축하고
다시 소리로 복원하는 신경망 코덱

핵심 스펙

- 24kHz 단일 채널 오디오
- 초당 50 프레임
- RVQ 12계층 사용
- 1초 → 600 토큰

Encoder: 24kHz 파형 → Acoustic Token
Decoder: Token → 24kHz 파형 복원

RVQ (Residual Vector Quantization) 12 계층

Layer 1	큰 특징 포착 (음정, 리듬)
Layer 2	1차 잔차(residual) 보정
Layer 3–4	세부 음색 추가
Layer 5–8	디테일 채움 (Coarse→Fine)
Layer 9–12	24kHz 고음질 마무리

1초 오디오 = 50 프레임 × 12계층 = 600 토큰 | 6kbps 초저비트레이트로 24kHz 고음질 복원 가능

세 모듈의 역할 분담

독립적으로 사전 학습 → MusicLM에서 결합

MuLan

조건부 토큰 (M)

- 텍스트↔오디오 공유 공간
- 128차원 임베딩
- RVQ 12계층 양자화
- 대조 학습(Contrastive)
- 분위기·스타일 조건 담당

w2v-BERT

의미 토큰 (S)

- MLM 자기지도 학습
- 600M 파라미터
- 1초 = 25 토큰
- k-means 1024 클러스터
- 멜로디·구조 정보 담당

SoundStream

음향 토큰 (A)

- 오디오 압축·복원 코덱
- RVQ 12계층 구조
- 1초 = 600 토큰
- 비트레이트 6kbps
- 음색·질감 정보 담당

↑ 세 모듈이 결합되어 MusicLM 계층적 생성의 입력으로 사용됩니다 ↑

Part 3

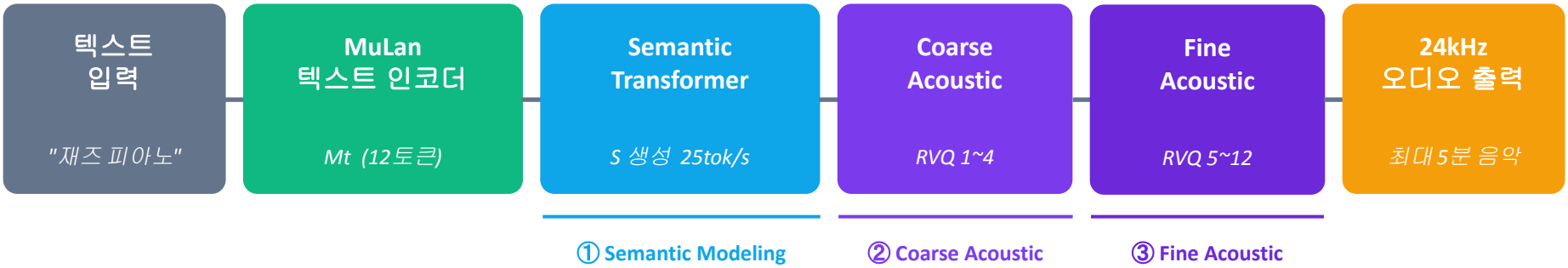
계층적 생성 파이프라인

왜 3단계 분업이 필요한가?

24kHz 오디오는 1초에 24,000개의 숫자 → AI가 한 번에 일관성 있게 생성하기 불가능
큰 그림(멜로디) → 음색 → 고음질 디테일 순서로 단계별 생성

전체 생성 파이프라인

텍스트 → MuLan 토큰 → Semantic → Coarse Acoustic → Fine Acoustic → 오디오



① Semantic Modeling

w2v-BERT 기반 장기 구조·멜로디
Mt 조건으로 S를 AR 방식으로 예측
초당 25 토큰 · 10초 → 250 토큰

② Coarse Acoustic

SoundStream RVQ 1~4계층 예측
기본 음색·질감을 씌움
S + Mt 조건 · 10초 → 2000 토큰

③ Fine Acoustic

SoundStream RVQ 5~12계층 예측
노이즈 제거·24kHz 고음질 완성
Coarse A 조건 · 10초 → 4000 토큰

Step ① Semantic Modeling

MuLan 조건 하에 자기회귀(AR) 방식으로 의미 토큰 시퀀스 생성

$$P(St \mid S<t, Ma)$$

St	현재 시점 t 의 의미 토큰
S<t	이전까지 생성된 모든 의미 토큰
Ma	MuLan 오디오 (또는 텍스트) 임베딩

작동 원리 (자기회귀 생성 AR)

① 텍스트를 MuLan 텍스트 인코더로 변환 → 12개의 조건 토큰 Mt 획득
② Semantic Transformer가 Mt를 보고 첫 S_1을 확률적으로 샘플링
③ S_1 포함 전체 맥락을 보고 S_2 예측 → S_3, S_4, ... 반복 (AR)
④ 10초 분량 = 초당 25개 × 10초 = **250개 Semantic Tokens** 생성

Transformer 스펙: 430M 파라미터 · 24 Layer · 임베딩 1024차원 · 드롭아웃 0.1

Step ②③ Acoustic Modeling

Semantic Token을 조건으로 실제 음향 특성 생성

② Coarse Acoustic Modeling

예측 대상: RVQ 1~4 계층

조건: Mt + Semantic Tokens(S)

역할: 기본 음색과 질감을 씌움

수식:

$$P(A1{:}4 \mid S, Ma)$$

10초 기준:
50 프레임 × 4계층 × 10초=
2000 토큰

③ Fine Acoustic Modeling

예측 대상: RVQ 5~12 계층

조건: Coarse Acoustic (A1:4)

역할: 노이즈 제거, 24kHz 고음질 완성

수식:

$$P(A5{:}12 \mid A1{:}4)$$

10초 기준:
50 프레임 × 8계층 × 10초=
4000 토큰

Part 3

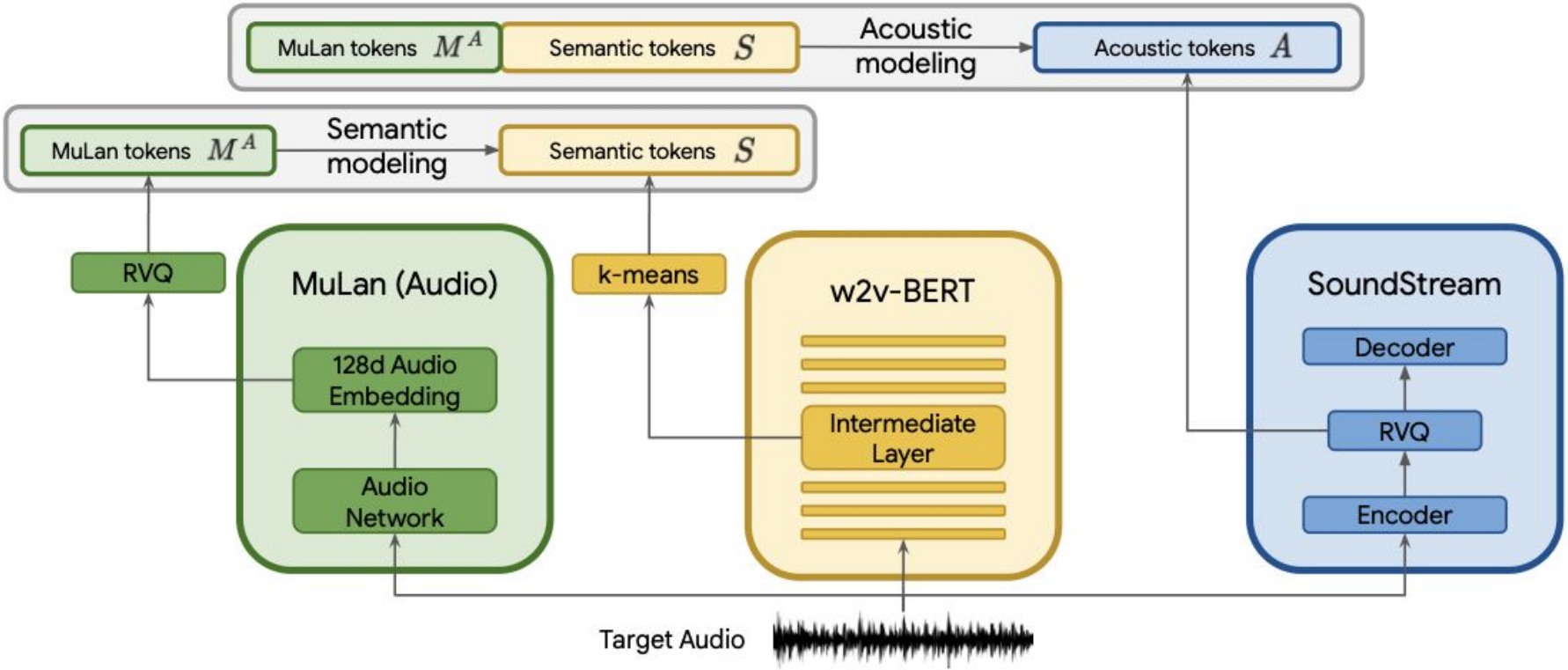
전체 생성 파이프라인

Training vs Inference

학습 → 24kHz 오디오
실제 음악 생성 → 사용자 프롬프트

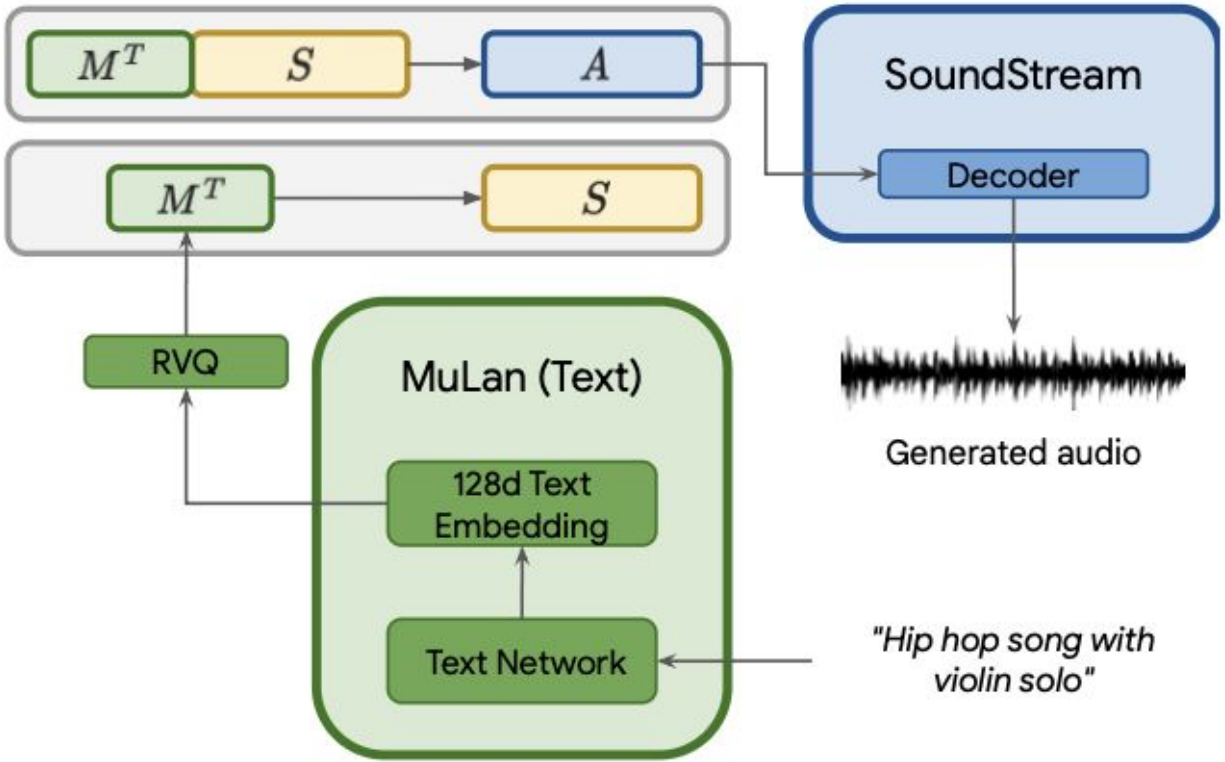
전체 생성 파이프라인 (Training)

정답지 토큰 생성 → Transformer 학습



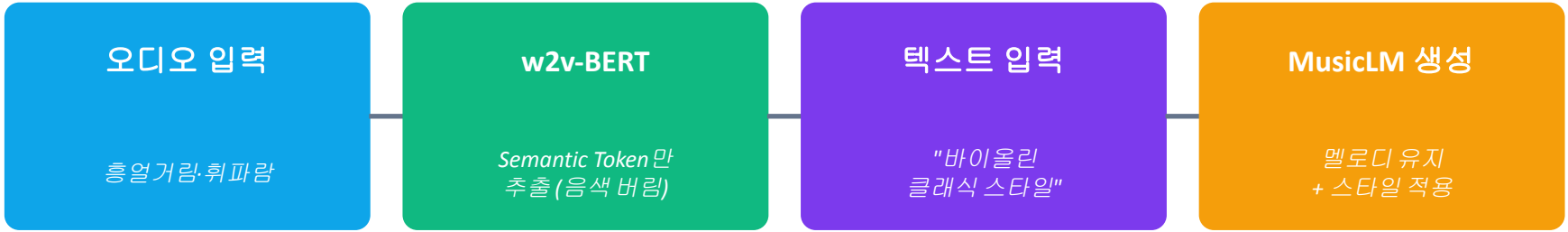
전체 생성 파이프라인 (Inference)

텍스트 → MuLan 토큰 → Semantic → Coarse Acoustic → Fine Acoustic → 오디오



멜로디 컨디셔닝 (Melody Conditioning)

흥얼거림·휘파람으로 멜로디를 유지하며 스타일 변환



핵심 아이디어

Semantic Token = 음악의 뼈대 (멜로디 + 리듬 + 구조) → 음색 정보는 포함하지 않음

같은 Semantic Token에 다른 음향 조건을 씌우면 멜로디는 같지만 스타일이 다른 음악이 생성됩니다

Part 4

실험 & 결과

Experimental Setup · Metrics · Quantitative Results

실험 설정 (Experimental Setup)

학습 데이터

- 500만 개의 오디오 클립 (~28만 시간, 24kHz)
- 텍스트 없이 음악 오디오만 사용

모델 스펙 (각 단계 동일)

- Decoder-only Transformer · 파라미터 430M
- 24 Layer
- 임베딩 1024차원 · 피드포워드 4096 · 드롭아웃 0.1

평가 데이터셋 : MusicCaps

AudioSet 기반 · 5,500개 음악 클립 · 10명 전문 음악가 영문 캡션 (평균 4문장) · 장르·무드·템포·악기 등 평균 11개 속성 포함

추론(Inference): 오디오 토큰 Ma 대신 텍스트 토큰 Mt 사용

평가 지표 (Evaluation Metrics)

FAD

Fréchet Audio Distance

$$FAD = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

μ_r, μ_g: 실제 오디오와 생성된 오디오 임베딩의 **평균 벡터** (전반적인 소리의 특징 차이)

Σ_r, Σ_g: 두 임베딩의 **공분산 행렬** (소리의 다양성 및 패턴 차이)

Tr: 행렬의 대각합(Trace)

↓ 낮을수록 실제 오디오에 가까움

평가 지표 (Evaluation Metrics)

KLD

Kullback-Leibler Divergence

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

$P(x)$: 실제 기준(Reference) 음악이 클래스 x(예: 드럼, 피아노)를 가질 확률

$Q(x)$: 생성된 음악이 클래스 x를 가질 확률

↓ 낮을수록 원본과 비슷한 특성

평가 지표 (Evaluation Metrics)

MCC

MuLan Cycle Consistency

$$MCC = \frac{v_T \cdot v_A}{||v_T||||v_A||}$$

$v_T \cdot v_A$ **(내적, Dot Product)**

$||v_T||||v_A||$ **(벡터 크기의 곱으로 나누기)**

↑ 높을수록 텍스트와 음악 잘 매칭

정량적 결과 비교

MusicLM vs Mubert vs Riffusion — 모든 지표에서 우수

MODEL	FAD_{TRILL} ↓	FAD_{VGG} ↓	KLD ↓	MCC ↑	WINS ↑
RIFFUSION	0.76	13.4	1.19	0.34	158
MUBERT	0.45	9.6	1.58	0.32	97
MUSICLM	0.44	4.0	1.01	0.51	312
MUSICCAPS	-	-	-	-	472

Ablation Study: Semantic Token을 제거하면 ?

설정: Mt → SoundStream 직접 연결
(Semantic Token 없이 바로 Acoustic 생성)

결과: 음질은 유사하지만 KLD↑ MCC↓
→ 장기 구조 붕괴 + 텍스트 일치도 저하

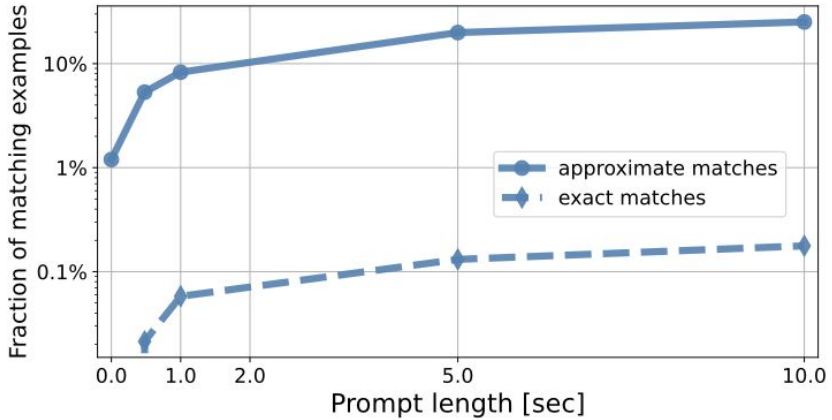
표절 검증 & 결론

🔍 표절 검증

방법: 최적 운송(Optimal Transport) 알고리즘으로 분포 간 이동 비용 계산

결과: 학습 데이터와 완전히 동일한 생성 결과
0.2% 미만

→ *암기(memorization)가 아닌 창작(generation)을 합니다*



MusicLM 핵심 기여 요약

① MuLan으로 28만 시간 무레이블 오디오를 학습 데이터로 활용 ② 3단계 계층적 생성으로 24kHz 고음질 확보
③ 멜로디 컨디셔닝으로 크리에이티브 제어 확장

감사합니다

QnA

MusicLM: Generating Music From Text

Agostinelli et al., Google Research · 2023

Dataset: MusicCaps · kaggle.com/datasets/googleai/musiccaps