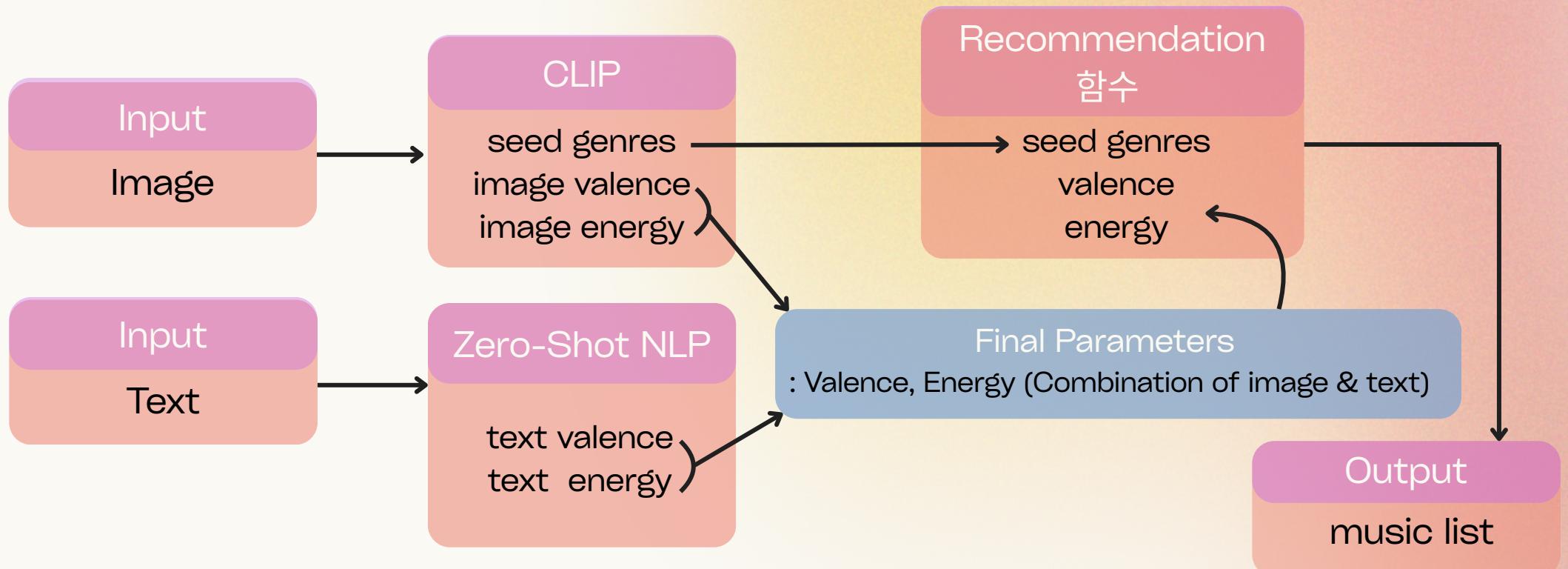


# AI DJ

## : Curating the Moment

Translating **Visual Context** and **Verbal Intent** into Spotify Audio Parameters



### • Background

이 프로젝트는 인스타그램, 유튜브 등 상황과 감정을 공유하는 소셜 미디어 환경에서 출발했습니다. 오늘날 사람들은 사진과 음악을 통해 자신의 상황과 감정을 표현하려 하고, 이 과정에서 상황과 감정에 어울리는 음악을 찾고 싶어합니다.

하지만, 기존 음악 추천 시스템에는 한계점이 있습니다. 먼저, 대부분의 추천 시스템은 사용자의 과거 청취 이력이나 선호도에 기반하여 음악을 추천합니다. 하지만, 이러한 시스템은 사용자의 순간의 감정을 반영하지 못합니다. 또한, 보편적인 상황과 사용자의 실제 감정이 다를 수 있습니다. 예를 들어, 사람은 비가 오거나 어두운 분위기 속에서 편안함이나 긍정적인 감정을 느끼기도 합니다. 하지만, 기존 시스템은 이러한 감정의 연속적인 스펙트럼을 포착하지 못합니다.

따라서, 저희 프로젝트는 이러한 문제를 해결하고 개인화된 실시간 감성 추천을 제공하는 것을 목표로 합니다.

2. 높은 유사도를 가진 카테고리를 결정합니다. 결정된 카테고리에 매핑된 기본 Spotify 파라미터(target\_energy, target\_valence, seed\_genres)가 Base Parameter으로 설정됩니다.
3. **텍스트 분석:** 입력된 텍스트는 Zero-Shot Classification에 의해 분석됩니다. 이는 텍스트가 감정(Valence: 긍/부정)과 에너지 (Energy: 활기/차분)에서 어느 방향을 지향하는지, 그리고 그 표현에 대한 확신도(Confidence)를 추출합니다.
4. **파라미터 튜닝:** 이미지를 베이스로, 텍스트를 가중치로 사용하여 파라미터를 정교하게 수정합니다. **Final Parameter = Base Parameter ± (최대 변동폭 × Confidence)**
5. **Spotify API 통합과 결과 출력:** 최종적으로 계산된 파라미터는 Spotify의 추천 함수에 적용됩니다. 이미지에서 추출한 시드 장르와 최종 파라미터에서 추출된 valence, energy 값이 Spotify의 recommendation 함수에 적용됩니다. 이 결과, 사용자의 상황, 감정에 적합한 음악을 제공합니다.

### • System architecture

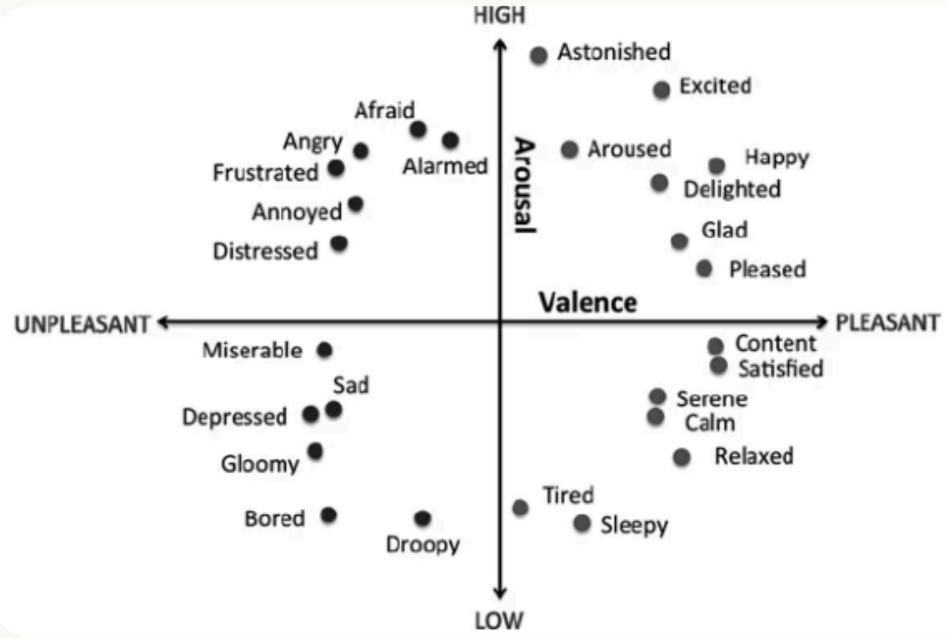
이 시스템은 이미지의 배경과 텍스트의 의도를 융합하여 Spotify 음악 추천 파라미터를 동적으로 미세조정 하는 멀티모달 시스템입니다.

1. **사용자 입력:** 사용자는 자신이 원하는 이미지를 제공하고, 그 상황에서 느끼는 감정과 의도를 텍스트로 입력합니다.
2. **이미지 분석:** 입력된 이미지는 CLIP(Contrastive Language-Image Pre-training)에 의해 분석됩니다. CLIP은 이미지를 사전 정의된 12가지 카테고리 중 가장

### • Data Definition

1. Spotify Audio Features - (0.0 ~ 1.0)  
**Valence, Energy, Acousticness, Danceability, Instrumentalness, Tempo, Key, Loudness**
2. Our Selection: **2-Dimensions-Russell's Model based**
  - (1) **Valence** (x축: 긍정-부정)
    - Spotify 정의: 트랙이 전달하는 음악적 긍정성
    - '감정의 색깔(Mood)'을 대변.
  - (2) **Energy** (y축: 활기-차분)
    - Spotify 정의: 활동성과 강렬함
    - '감정의 강도(Intensity)'를 대변.

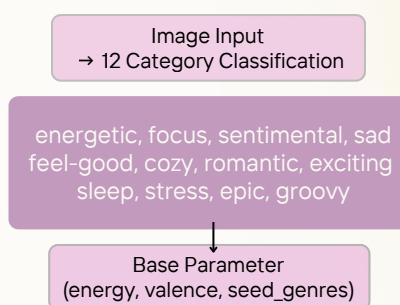
## Russell's circumplex model



### • Image Analysis: CLIP

CLIP 모델은 Open AI에서 개발한 멀티모달 AI 모델로 이미지와 텍스트를 함께 학습하여, 하나의 공간에서 매핑합니다.

본 프로젝트에서 CLIP 모델은 사용자 입력 이미지를 분석하여 12가지로 미리 정의된 감성 카테고리 중 가장 적합한 카테고리를 찾는데 사용되었습니다.



### • Scenario example

#### CLIP 이미지 분석

- 상황: [Sentimental, melancholic rainy day]
- 카테고리 결정 신뢰도 : 96.3%
- BaseParameter : V : 0.25 / E : 0.3



### • Text Analysis: Zero-Shot NLP

mDeBERTa-v3-base-mnli-xnli

일반적인 분류가 아닌, 자연어 추론(Natural Language Inference, NLI) 방식을 사용하여 텍스트를 분석합니다.

#### Step 1. 가설 생성

사용자 입력(X)과 분석하고 싶은 라벨(L)을 결합한 가설 문장 생성,

- X: “비가 오지만, 오늘 좋은 일이 있어서 신나!”
- L1: 긍정(+) / L2: 중립 / L3: 부정(-)
- H1: “이 텍스트는 긍정적인 내용이다.”
- H2: “이 텍스트는 중립적인 내용이다.”
- H3: “이 텍스트는 부정적인 내용이다.”

#### Step 2. 추론 및 확률값 산출

모델은 각 쌍(X, H)에 대해 참일 가능성을 나타내는 확률값 출력.

- (X, H1) : 0.93 / (X, H2) : 0.02 / (X, H3) : 0.03

#### Step 3. 최종 출력

가장 높은 확률값 선택 및 가중치로 사용.

- (X, H1) : 0.93 > Confidence = 0.93

### • Scenario example

#### Zero-Shot NLP 분석

- H: “나가서 비를 맞으며 신나게 놀고 싶은 기분인걸!”
- V축 분석 : 긍정, 신뢰도 : 0.858
- E축 분석 : 활기찬, 신뢰도 : 0.993

### • Parameter Tuning

#### 1. 튜닝 알고리즘

$$P_{final} = P_{base} + (Direction \times \Delta_{max} \times Confidence)$$

- $P_{final}$ : 최종 추천에 사용될 Spotify 파라미터 (Energy, Valence)
- $P_{base}$ : 이미지 분석을 통해 얻은 기준값
- Direction: 감정의 방향 (+1: 긍정/활기, -1: 부정/차분)
- $\Delta_{max}$ : 설계된 최대 변화폭 상수 (0.4)
- Confidence: NLP 모델이 추출한 감정의 신뢰도 (0.0 ~ 1.0)

#### 2. 설계논리

##### • $P_{base}$ (Base Parameter) : "The Anchor"

**How:** 12개 감성 카테고리별로 Spotify의 대표 트랙 1,200곡(각 100곡)의 오디오 특성(Audio Features) 평균값을 추출하여 DB화

**Role:** 텍스트가 아무리 변해도 음악이 이미지의 기본 맥락(Context)을 벗어나지 않도록 잡아주는 **닻(Anchor)** 역할.

##### • $\Delta_{max}$ (Max Delta) : "Safety Lock"

**Value:** 0.5 (Spotify 파라미터 범위 0.0~1.0 기준)

##### Why 0.5?:

텍스트가 긍정적이라고 해서 비 오는 날(Valence 0.2)에 댄스곡(Valence 0.9)을 틀면 **인지적 부조화(Cognitive Dissonance)**가 발생함. 최대 변화폭을 0.5로 제한하여, “비 오는 날의 무드”를 해치지 않는 선에서 가장 밝은 곡(Valence 0.6)까지만 추천되도록 **맥락 보존(Context Preservation)** 장치를 마련.

#### • Scenario example

H : “나가서 비를 맞으며 신나게 놀고 싶은 기분인걸!”

$$P_{final\_V} = 0.25 + (0.5 * 0.858) = 0.68$$

$$P_{final\_E} = 0.3 + (0.5 * 0.993) = 0.80$$

#### • Scenario example (comparison)

H : “비가 와서 습하지만.. 뭐, 나쁘지 않아”

$$P_{final\_V} = 0.25 + (0.5 * 0.553) = 0.53$$

$$P_{final\_E} = 0.3 + (0.5 * 0.737) = 0.67$$

### • Interpretation

같은 상황(비 오는 어두운 밤)에서도 사용자에 따라 느끼는 감정, 그 감정의 강도는 다릅니다. 사용자의 텍스트 입력, 즉 감정과 그 강도에 따라 파라미터를 섬세하게 조절함으로써 사용자가 가장 원하는 음악을 추천할 수 있도록 파라미터를 조절할 수 있게 설계하였습니다.

### • 시사점 및 확장성

본 시스템은 정성적 감성 데이터를 Spotify의 정량적 파라미터로 직접 연결하는 멀티모달 융합을 보여줍니다. 특히, 텍스트 분석에서 추출되는 신뢰도를 가중치로 활용하여, 단순 긍/부정이 아닌 감정의 연속적인 스펙트럼을 추천에 반영합니다. 이는 기존 시스템이 해결하지 못했던 상황과 의도의 충돌을 해결하고, 사용자의 감성에 맞게 개인화된 추천 경험을 제공한다는 점에서 큰 의의가 있습니다.

이러한 ‘기준점+가중치 튜닝’ 논리는 음악을 넘어 다양한 콘텐츠 분야로 확장될 수 있습니다. 파라미터 튜닝을 다른 음향 속성까지 확장하고, 사용자의 정보를 학습하여 더욱 정밀한 시스템으로 확장될 수 있습니다. 더 나아가, 영상, 광고, 콘텐츠 등에 접목하여 범용적인 감성 기반 추천 시스템으로 활용될 수 있습니다.