

Make the memory

신재함 이현수 임형섭 정수민 현재민



T아카데미
빅데이터분석가 4기

TEAM 보고싶조

CONTENTS

01 — 주제 및 개요



02 — 모델링

03 — 개선과정

04 — 추모 : 뉴모리얼

05 — 사운드 플레이그라운드

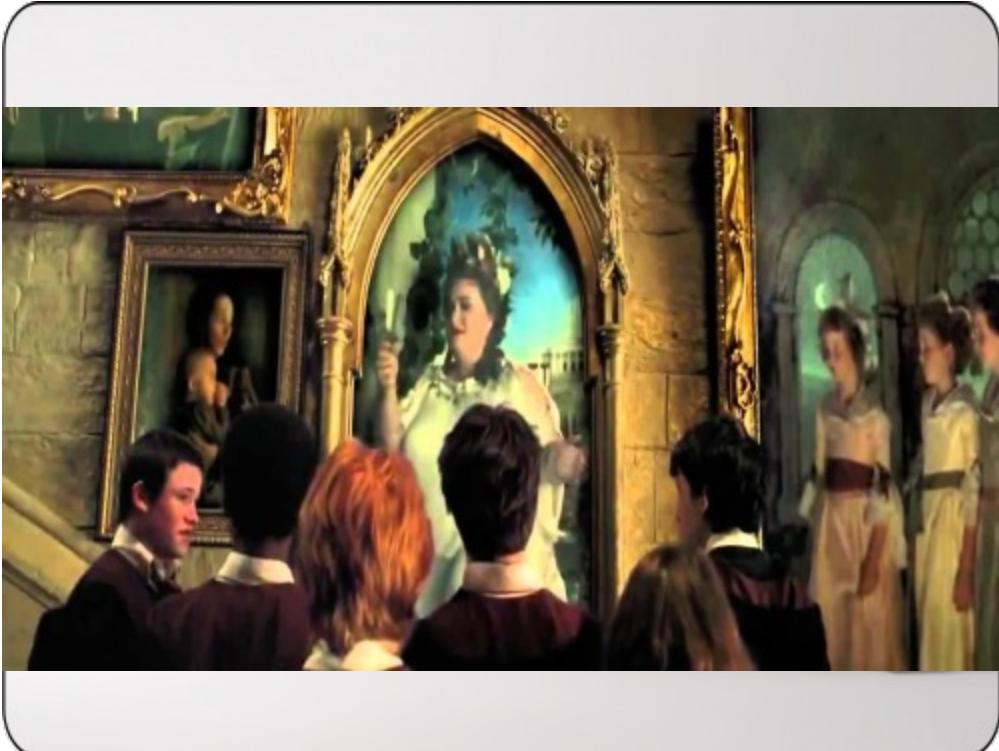
06 — 이모티콘 : 리얼티콘

07 — 가치 및 의의

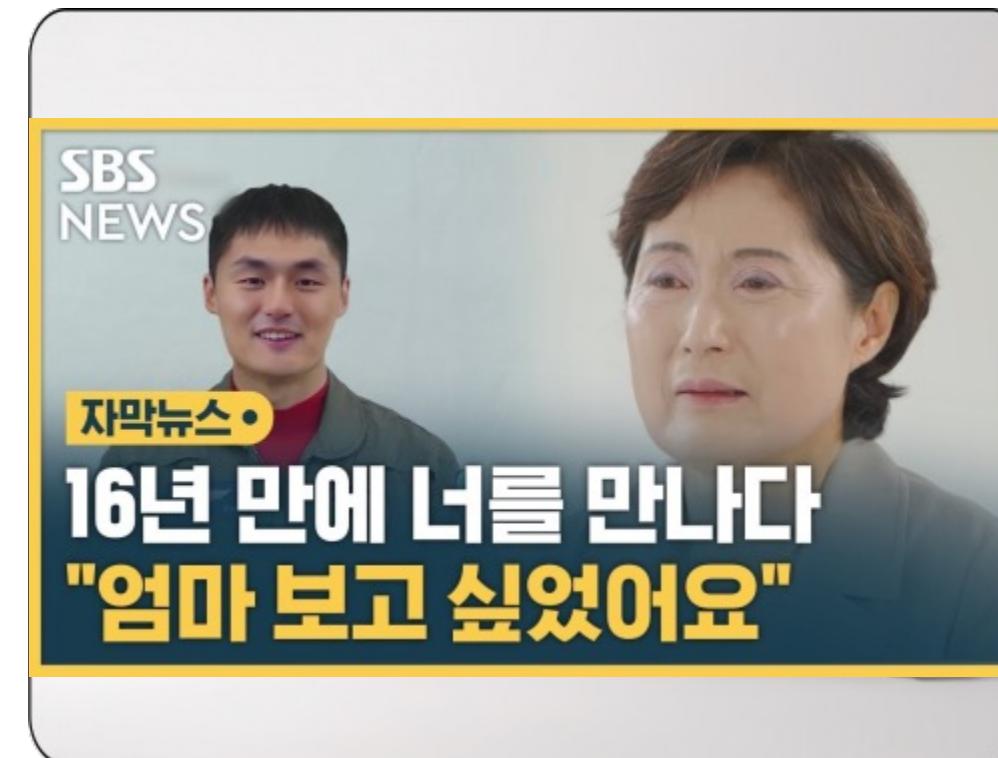


1. 주제 및 개요

유사 예시



해리포터 움직이는 액자



순직자 AI 복원



고 김광석 AI 노래 커버

유사 서비스



The screenshot shows the MyHeritage user interface. On the left, there's a "Gallery" section displaying a large black and white portrait of a young girl and several smaller circular thumbnails of other people. On the right, a modal window titled "Select animation" is open, divided into two sections: "General animations" (with numbered options #1 through #10) and "Special animations" (with icons for Smile, Dance 1, Dance 2, Kiss, Thankful, Kiss and wink, Eyebrows, Approval, Sideways, and Compassion). The "Kiss and wink" icon is highlighted with a red border.

저화질의 사진을 고화질로 변환
사진을 움직이는 사진으로 애니메이션화

→

한정적인 기능의 애니메이션
저렴하지 않은 비용

유사 서비스



Text to Video AI



[기능]

1. 텍스트 메세지를 비디오로 변환
2. 정적이미지의 애니메이션화
3. DALL·E 이미지 애니메이션
4. 비디오 길이를 앞뒤로 확장하여 편집 용이
5. 비디오-비디오 변환
6. 비디오 스티칭
7. 최대 2048×2048 해상도의 이미지 생성

유사 서비스

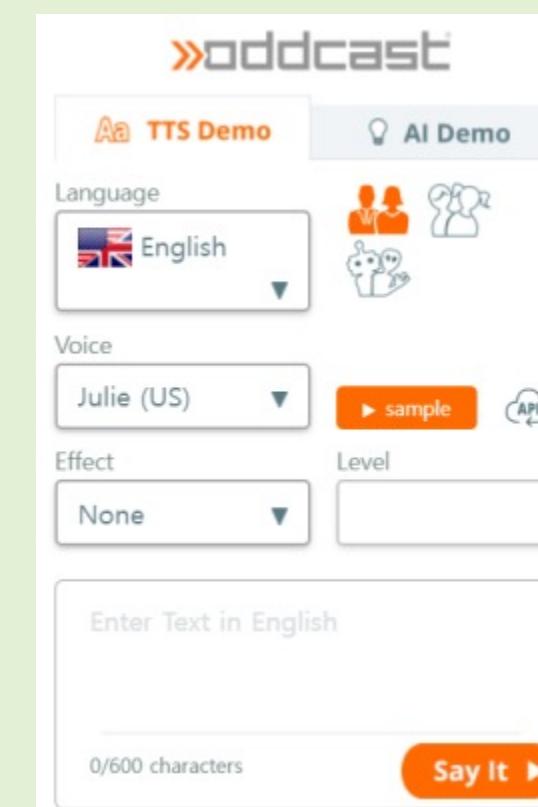
TextToSpeech

동영상 제작하면서 음성이 필요할 때, 팟캐스트를 만들 때,
특정 제품의 음성 소개가 필요할 때 사용하는 텍스트를 음성으로 변환하는 기능

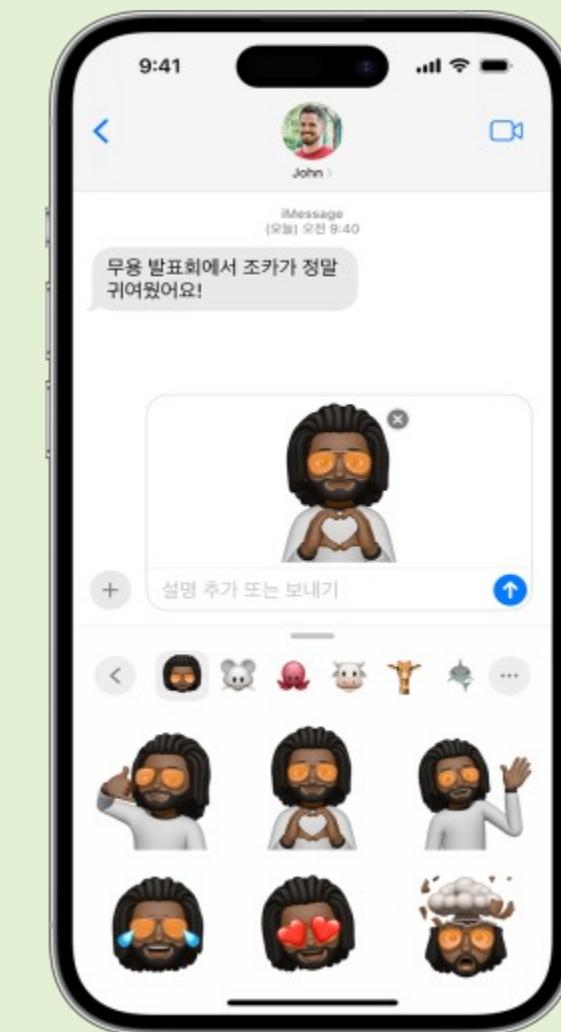
세상 모든 콘텐츠에 보이스를 더하다
CLOVA Dubbing^β



typecast
We invent the future of creativity with AI



MEMOJI STICKERS



보고싶조 Service



뉴모리얼

돌아가신 고인의 목소리를
복원하여 원하는 내용의
음성메세지 및 영상을 제공하는
AI 가상 추모 서비스



리얼티콘

사진 한 장으로
나만의 이모티콘을 제작할 수
있는 AI 서비스



사운드 플레이그라운드

5분 가량의 음성 파일만
있으면 내 목소리로
모든 노래를 부를 수 있다!?

Our Function

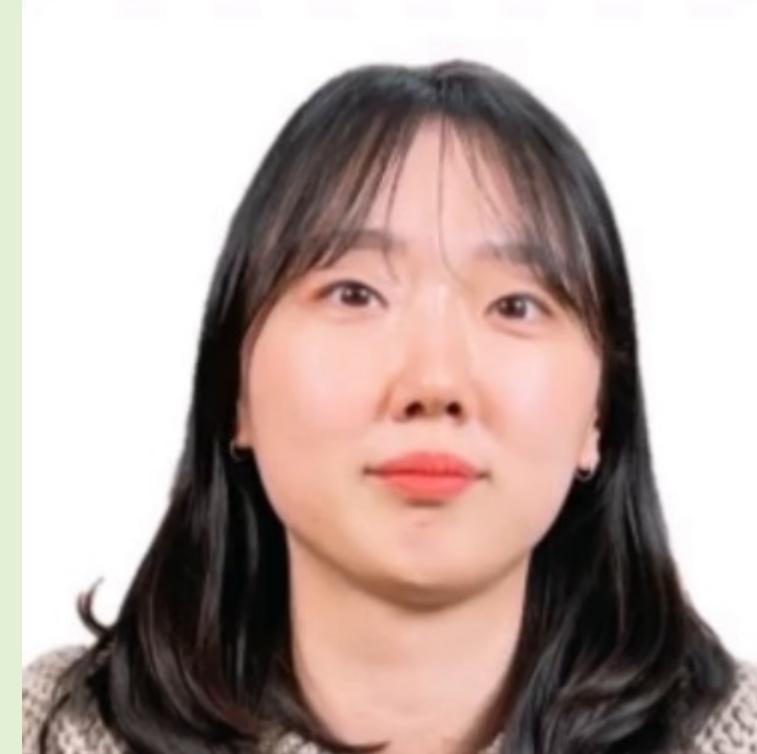
내 이미지로,



내 목소리를 사용해서,



나만의 비디오를 만듭니다.





2. 모델링

- 개발환경
- project overflow

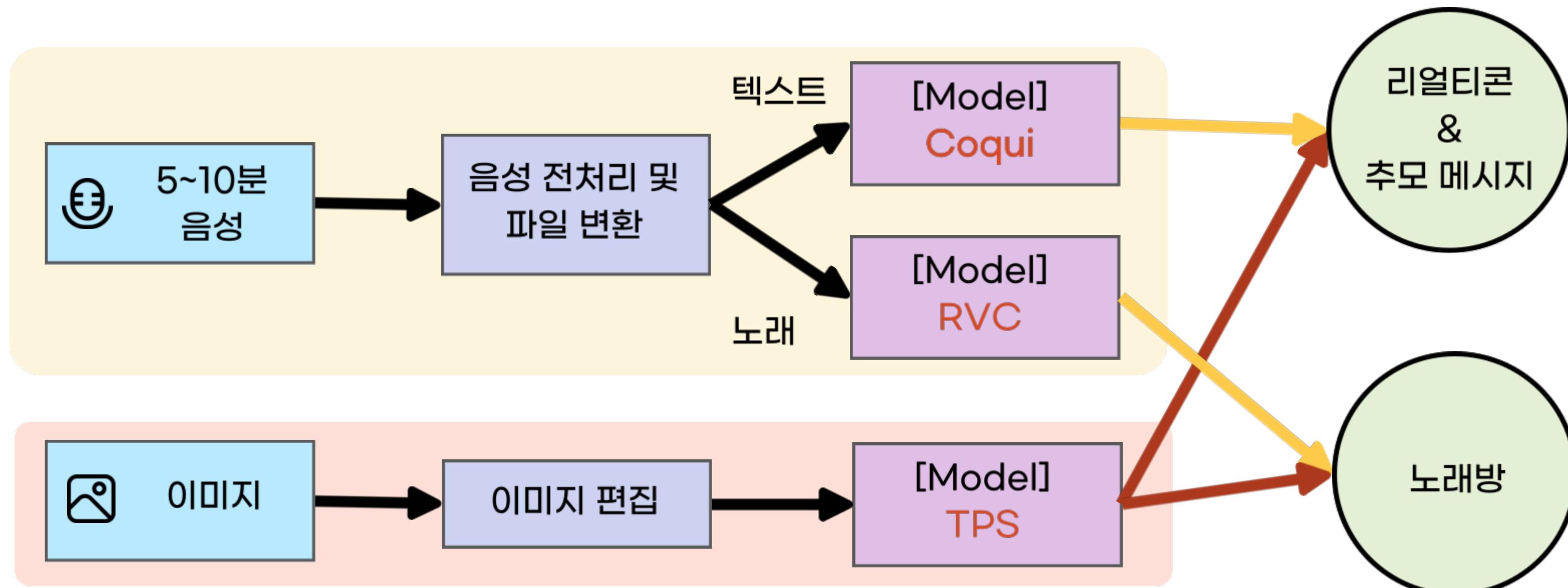
개발환경



Google Cloud Platform



Project Overflow



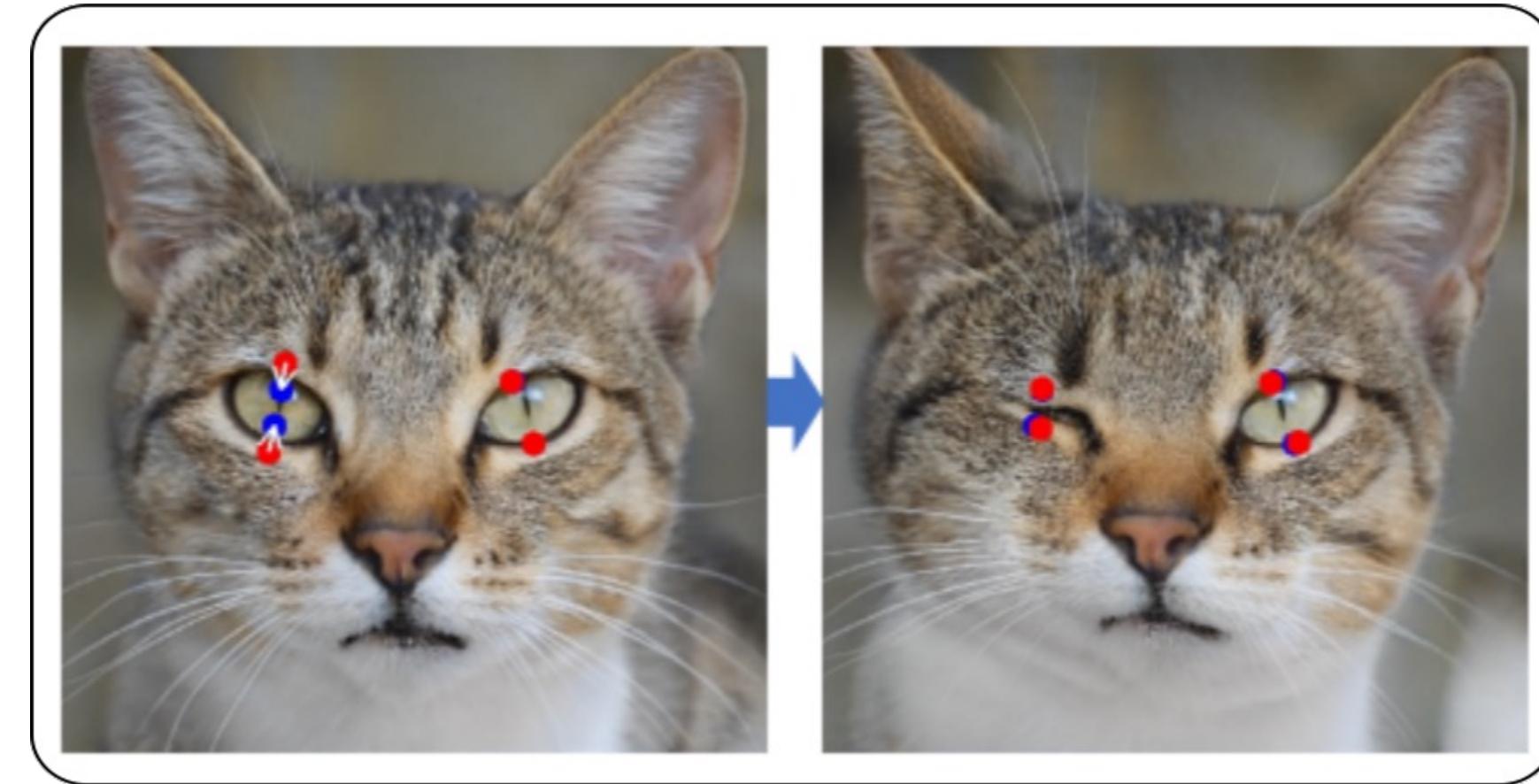
DragGAN

DragGAN

이전 모델인 GAN의 제어 가능성을 높여 이미지 개체의 유연성, 정밀도, 일반성을 종합적으로 다른 모델. 사용자 상호 작용 방식을 바탕으로 '드래그'하여 목표지점에 정확하게 도달하게 함

DragGAN만의 특징 2가지

1. 핸들포인트를 목표 위치로 이동하도록 구동하는 기능 기반 모션 감독
2. 핸들포인트를 추적하는 방식을 통해 각 step에서 핸들포인트의 위치를 지속적으로 파악



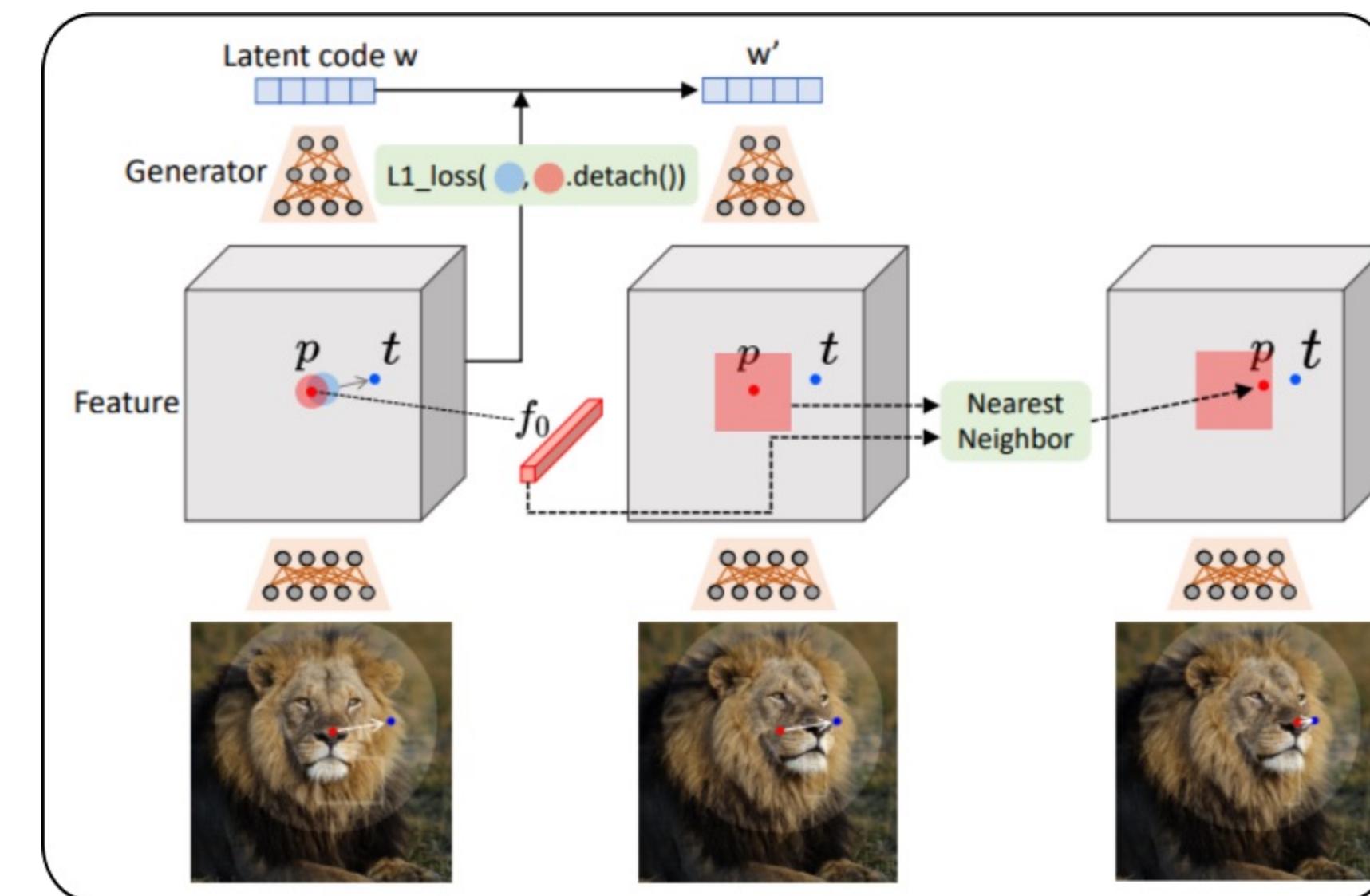
DragGAN

DragGAN

DragGAN을 통해 누구나 픽셀이 이동하는 위치를 정밀하게 제어하여 이미지를 변형할 수 있으며, 이를 통해 동물, 자동차, 인간, 풍경 등 다양한 카테고리의 포즈, 모양, 표정 및 레이아웃을 조작할 수 있음

한계: 드래그가 실시간으로 움직이는 것 때문에 실시간 적용에 부적합

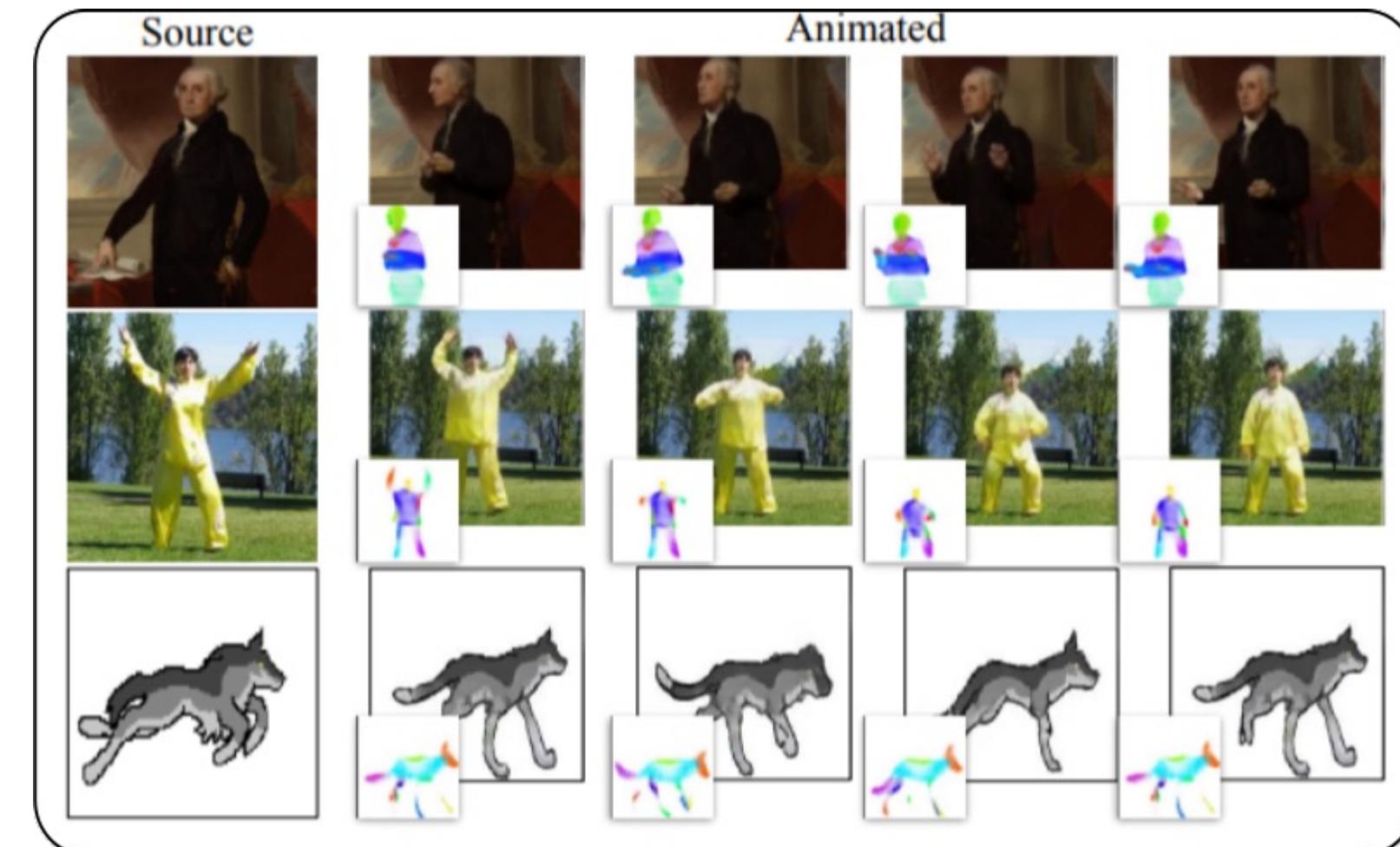
논문: Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold



Articulated-animation

Articulated-animation

- 해당 모델은 Region distribution과 stability를 개선했고, reconstruction accuracy와 유저 인식 성능이 좋았으며, 많은 region을 scale하는 능력 또한 특출났음
- 또한, TED-talks라는 굉장히 어려운 dataset을 제안
- domain data에 대해서는 약간의 결과를 냈지만, Inanimate object 까지 실용적으로 적용하기 위해서는 여전히 generalization은 어려운 문제로 남아있음.



Thin-Plate Spline Motion Model

Thin-Plate Spline Motion Model for Image Animation

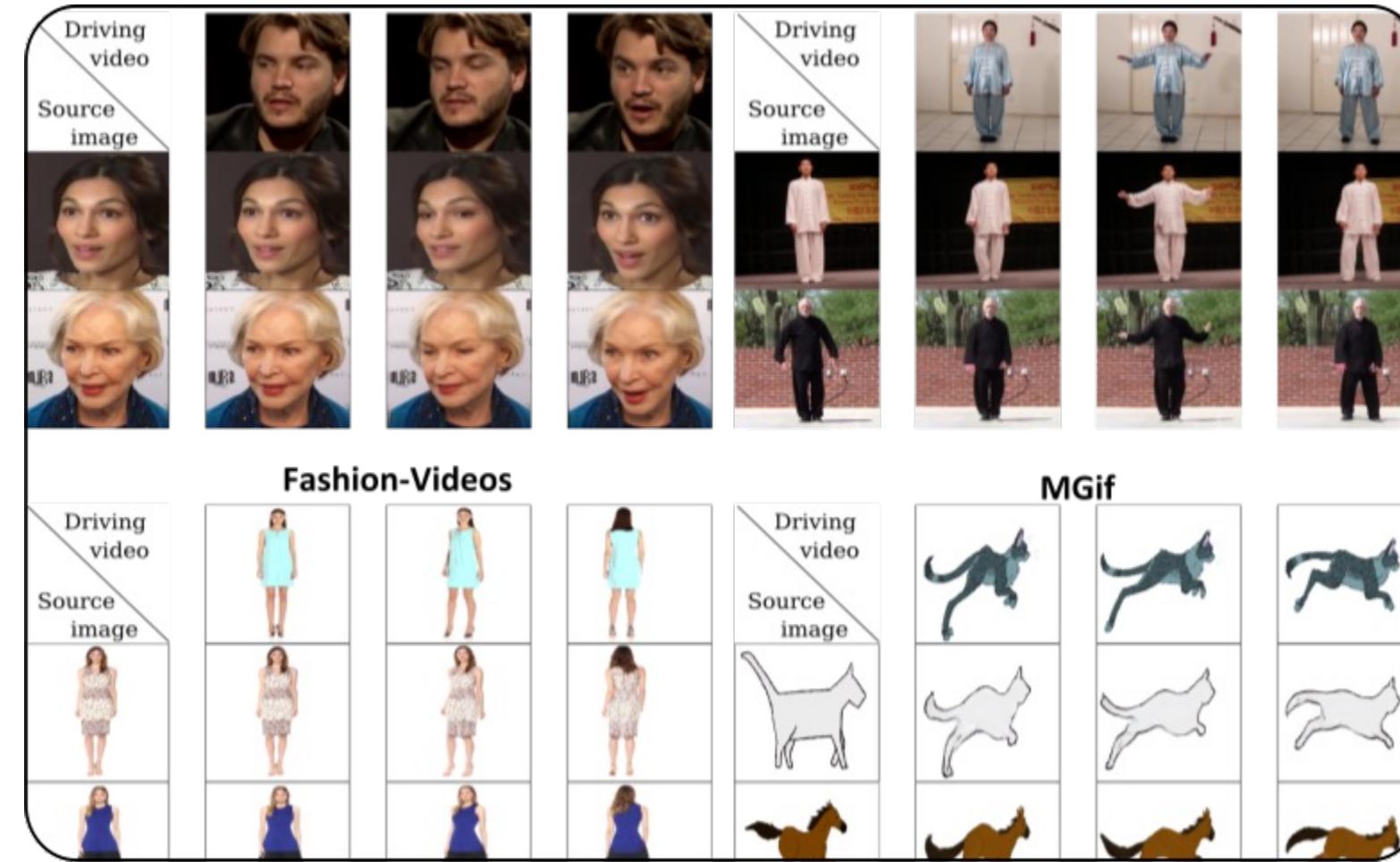


- 최근 연구는 사전 지식을 사용하지 않고 비지도 방법을 통해 임의의 개체에 대해 모션 전송을 수행하려고 시도함
- 그러나 원본 및 driving 이미지의 객체 사이에 큰 포즈 간격이 있을 때의 비지도 방법은 현재 여전히 중요한 과제로 남아 있음
- 본 논문에서는 이러한 문제를 극복하기 위해 새로운 end-to-end 비지도 모션 전송 프레임워크를 제안

Thin-Plate Spline Motion Model

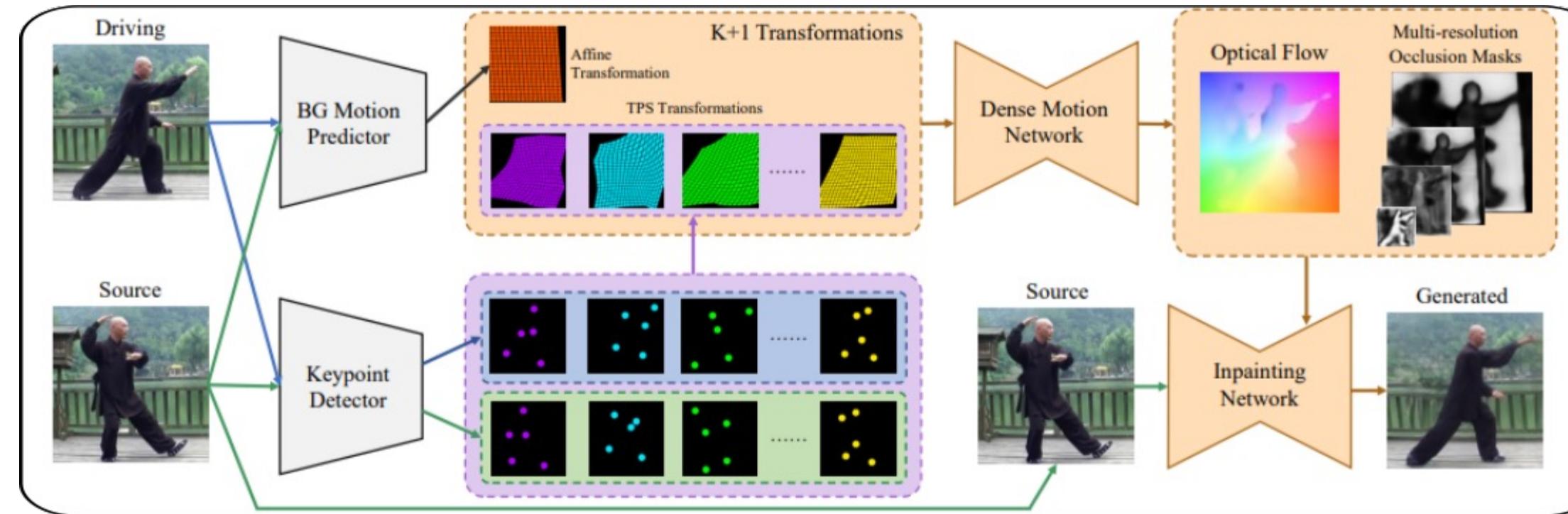
Thin-Plate Spline Motion Model for Image Animation

- 보다 유연한 optical flow를 생성하기 위해 thin-plate spline 모션 추정을 제안하는데, 이는 소스 이미지의 feature 맵을 주행 이미지의 feature 도메인으로 왜곡시킴
- 누락된 영역을 보다 현실적으로 복원하기 위해 다중 해상도 폐색 마스크를 활용하여 보다 효과적인 feature 융합을 달성
- 추가적인 보조 손실함수는 네트워크 모듈에 명확한 분업이 있도록 설계되어 네트워크가 고품질 이미지를 생성하도록 장려



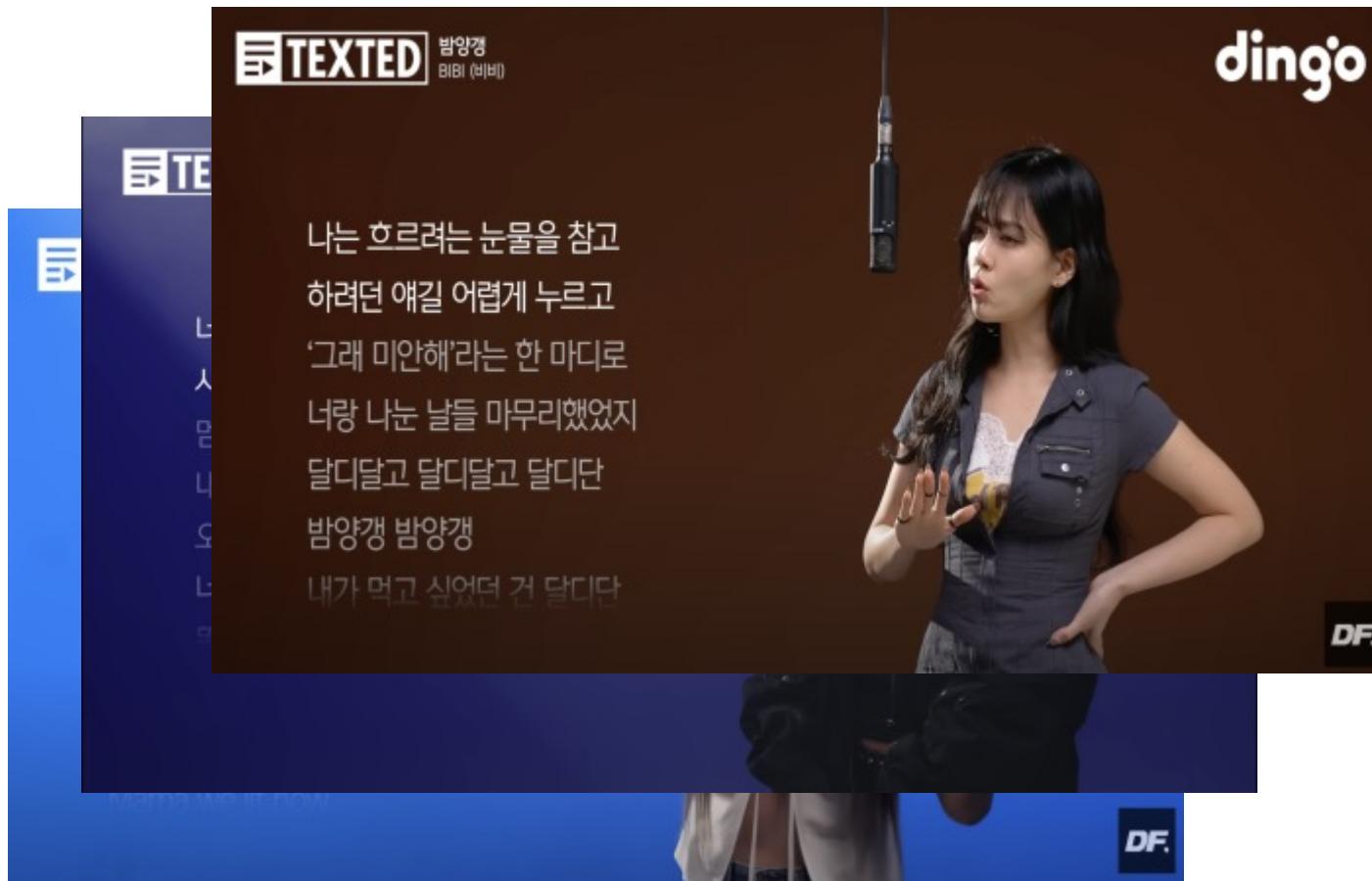
Thin-Plate Spline Motion Model

Thin-Plate Spline Motion Model for Image Animation



- 원본에서 driving 이미지까지의 움직임을 근사화하기 위해 TPS 움직임 추정 제시
- 훈련 초기 단계에서 여러 TPS 변환을 결합하기 전에 drop-out을 수행
- 새로운 end-to-end 비지도 모션 전송 프레임워크 제안 및 추정된 광학 흐름을 사용해 원본 이미지의 특징 맵을 왜곡
- 다중 해상도 폐색 마스크를 활용하여 인페인팅을 위해 누락된 영역을 나타냄
- 다양한 데이터셋에서 이전 비지도 모션 전송 방법보다 성능이 우수하며, 특히 모션 관련 지표가 눈에 띄게 개선된 모델

Thin-Plate Spline Motion Model



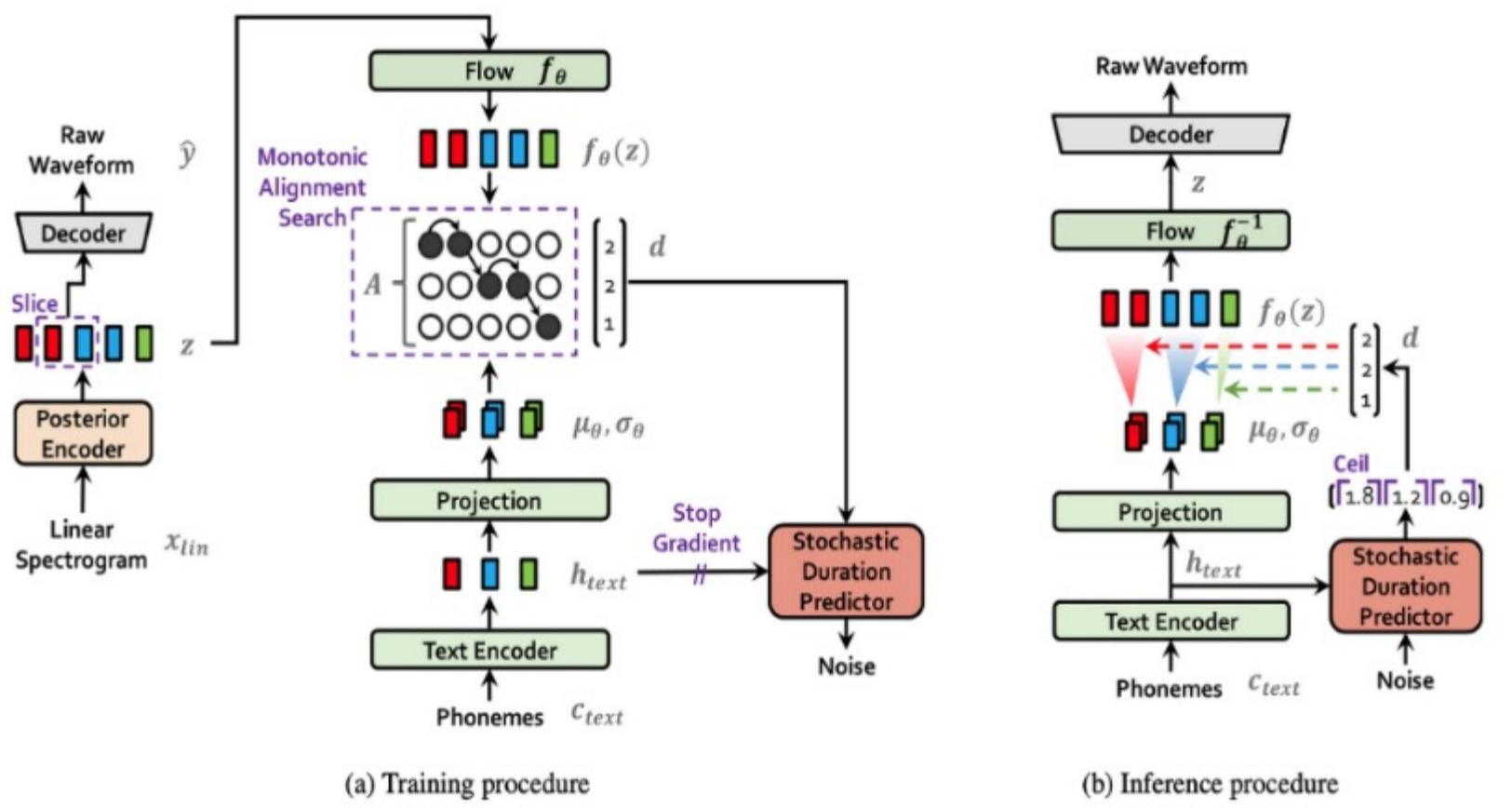
X 200

Thin-Plate Spline Motion Model



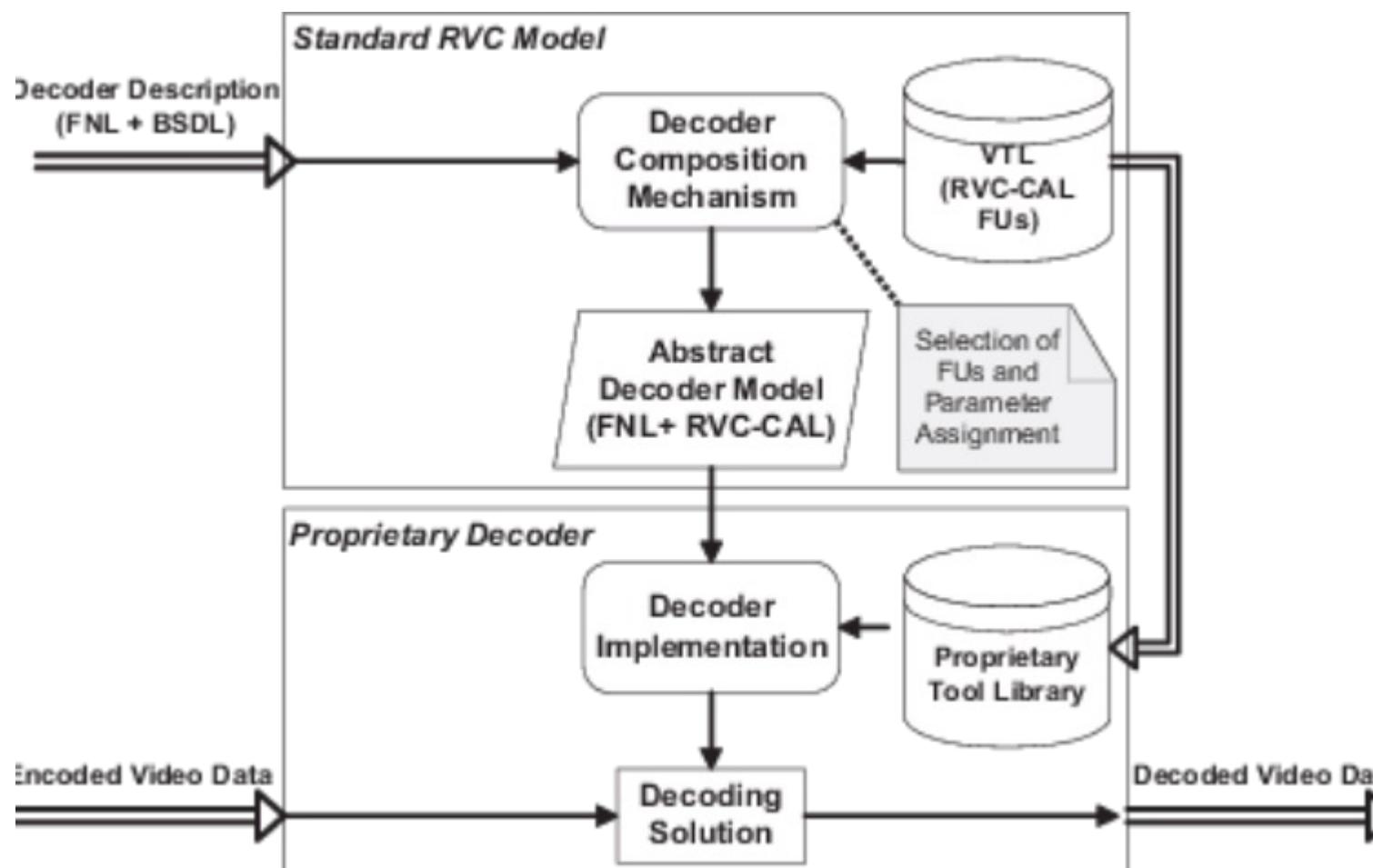
- 데이터의 라벨링이 필요없다는 비지도학습의 장점을 이용하여 사전학습된 모델인 VOX,TED를 제외한 저희 조 만의 학습 모델을 개발중
- 위의 이미지는 DINGO-TEXTED의 영상을 10초 단위로 자른 총 200개의 영상으로 진행중인 dingo 모델의 학습을 시각화 해서 보여준 이미지

Coqui - TTS 모델



- Tacotron, Tacotron2, Glow-TTS, SpeedySpeech을 사용한 딥러닝 모델
- Two-stage TTS 모델보다 더 자연스러운 음성을 생성하는 병렬 end-to-end TTS 모델
- spectrogram loss를 활용해 target과 생성된 음성 사이의 불일치를 완화
- Two-stage TTS는 병렬화가 어렵고, predefined intermediate feature로 인해 hidden representation도 활용하기 어려움
- 기존 TTS에 비해 병렬화가 가능하며, predefined intermediate feature로 인해 hidden representation도 활용하기 쉬움

TTS 모델 - RVC



Retrieval-based Voice Conversion

- 검색 기반 음성 변환
- AI 음성 합성 기술으로 기존의 Diff-SVC와 비슷한 형태이지만 Diff-SVC는 Stable Diffusion을 이용해 음파 이미지를 만드는 방식이고 RVC는 기존의 음성데이터를 이용해 변조를 하는 방식으로, 음성 변조와 비슷함



3. 개선과정

- 이미지개선
- 음성개선

Thin-Plate Spline Motion Model



Input



Output

1. 움직임이 크고, 배경이 복잡할 수록 배경이 심하게 일그러짐
=> 배경제거 / 단색 배경으로 진행
2. 안경의 빛 반사가 심하면 눈을 잘 인식하지 못함
=> 빛 반사를 최소화한 사진을 사용하거나 안경을 벗은 사진이 더 좋은 결과가 나옴

Thin-Plate Spline Motion Model

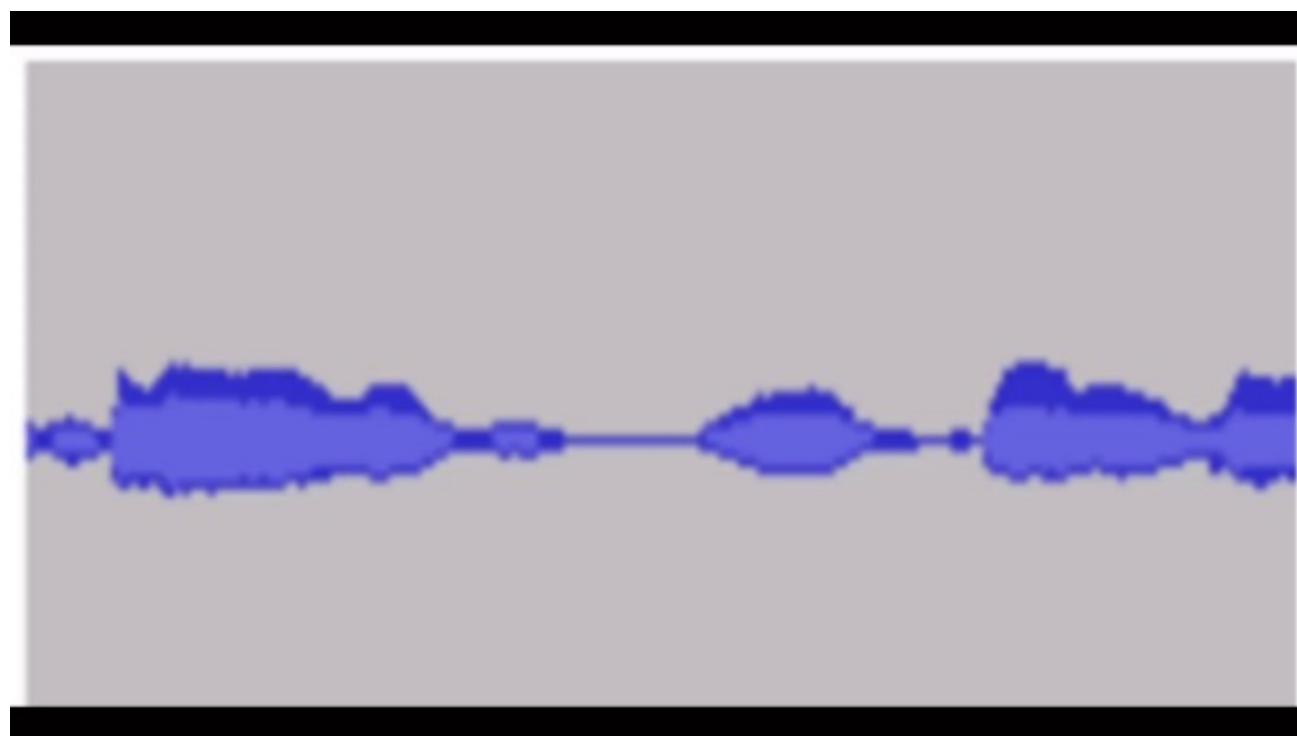
3. 영상 얼굴 사이즈가 작게 나오면 얼굴(입술 등)의 미세한 움직임을 사진에 적용이 잘 안됨
=> 영상의 얼굴 사이즈를 키워 적용



* 입을 작게 움직인 경우 인식을 못함

RVC 모델

데이터 음성 길이 - 3분

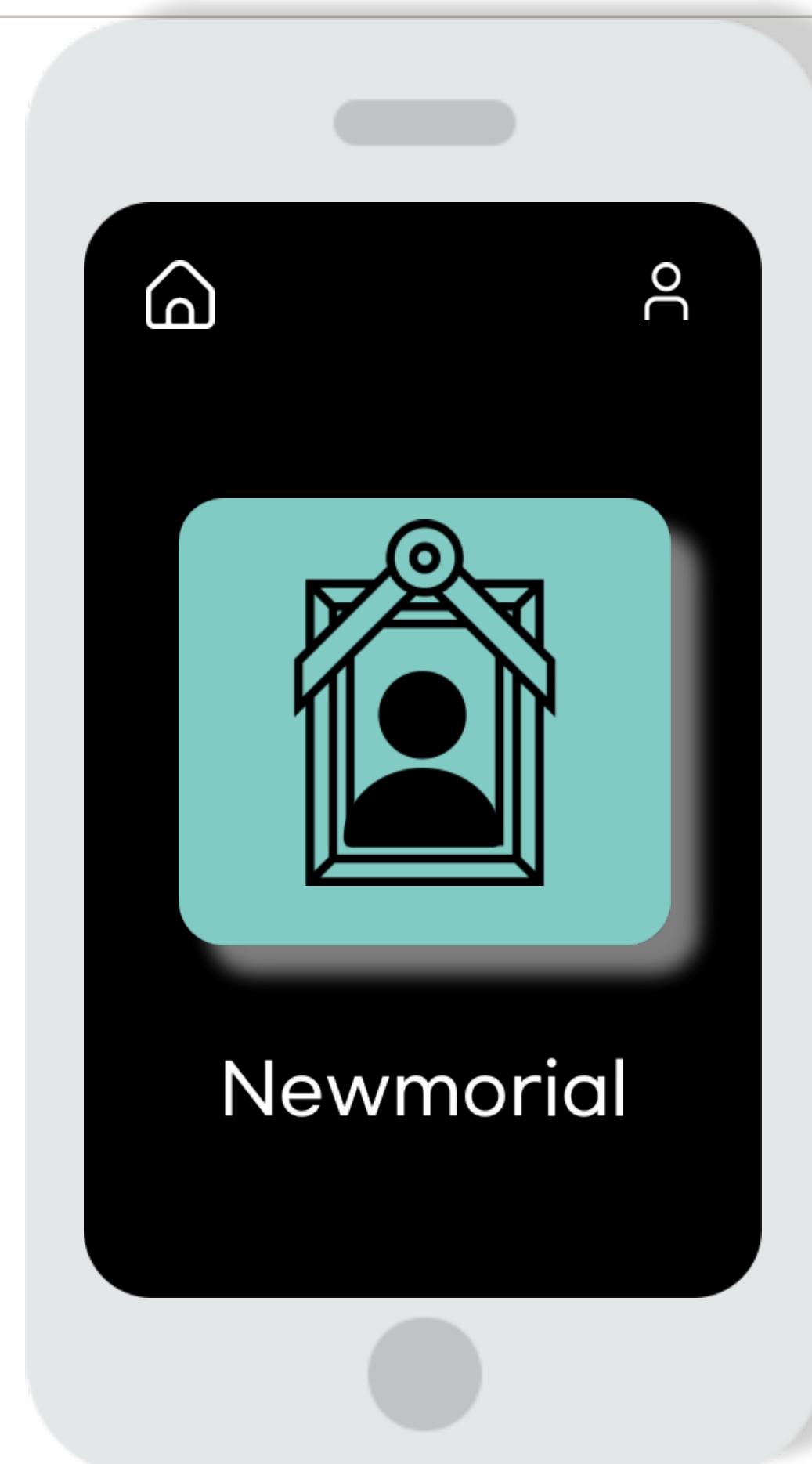


노이즈



데이터 음성 길이 - 7분





4. 뉴모리얼

(New + Memorial)

뉴모리얼 AI 가상음성+영상 추모 서비스



차별점

- 고인의 목소리를 복원하여 원하는 내용의 음성 메세지를 제공
- 고인의 사진을 실제 대화하는 모습과 표정으로 영상화



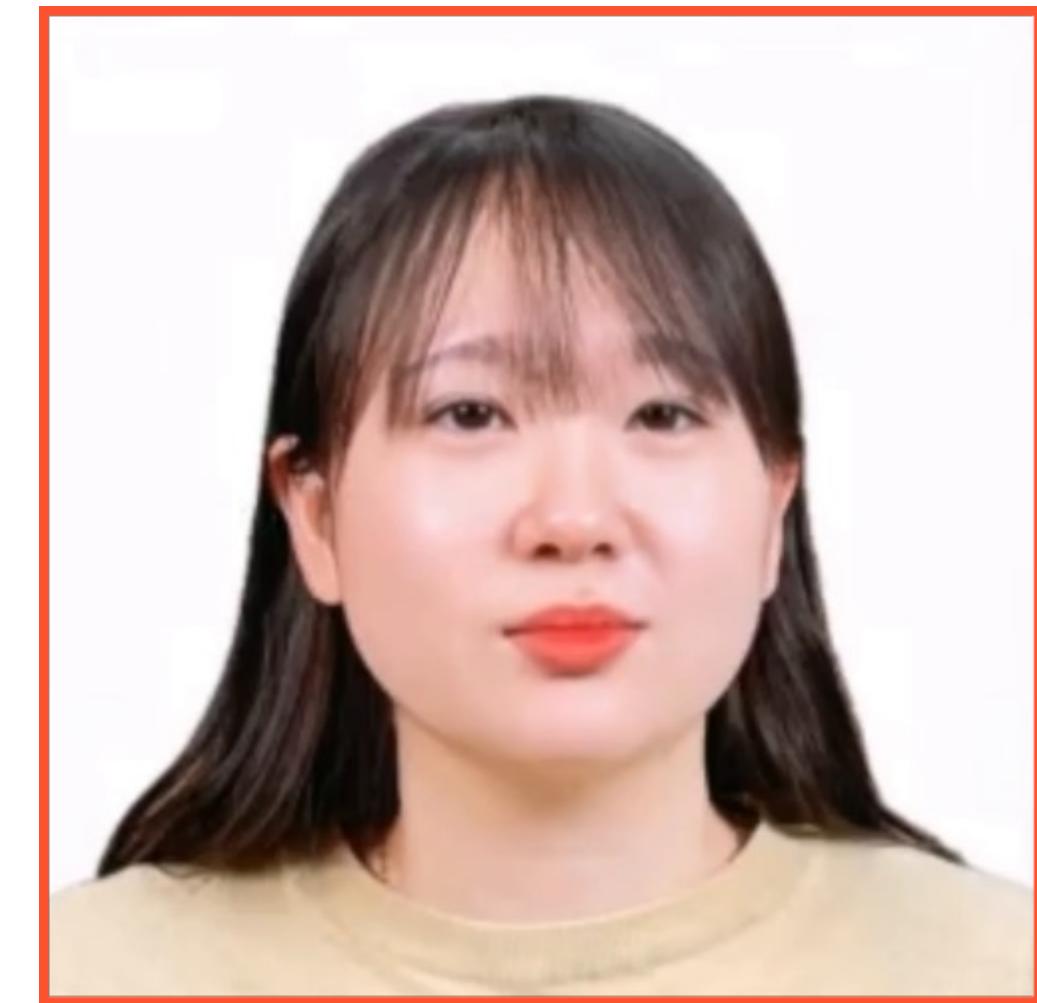
생성된 AI 음성 및 영상으로
"영상 메세지" 생성

뉴모리얼

생일축하 영상 서비스



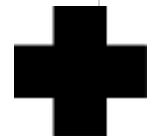
driving.mp4



Output

뉴모리얼

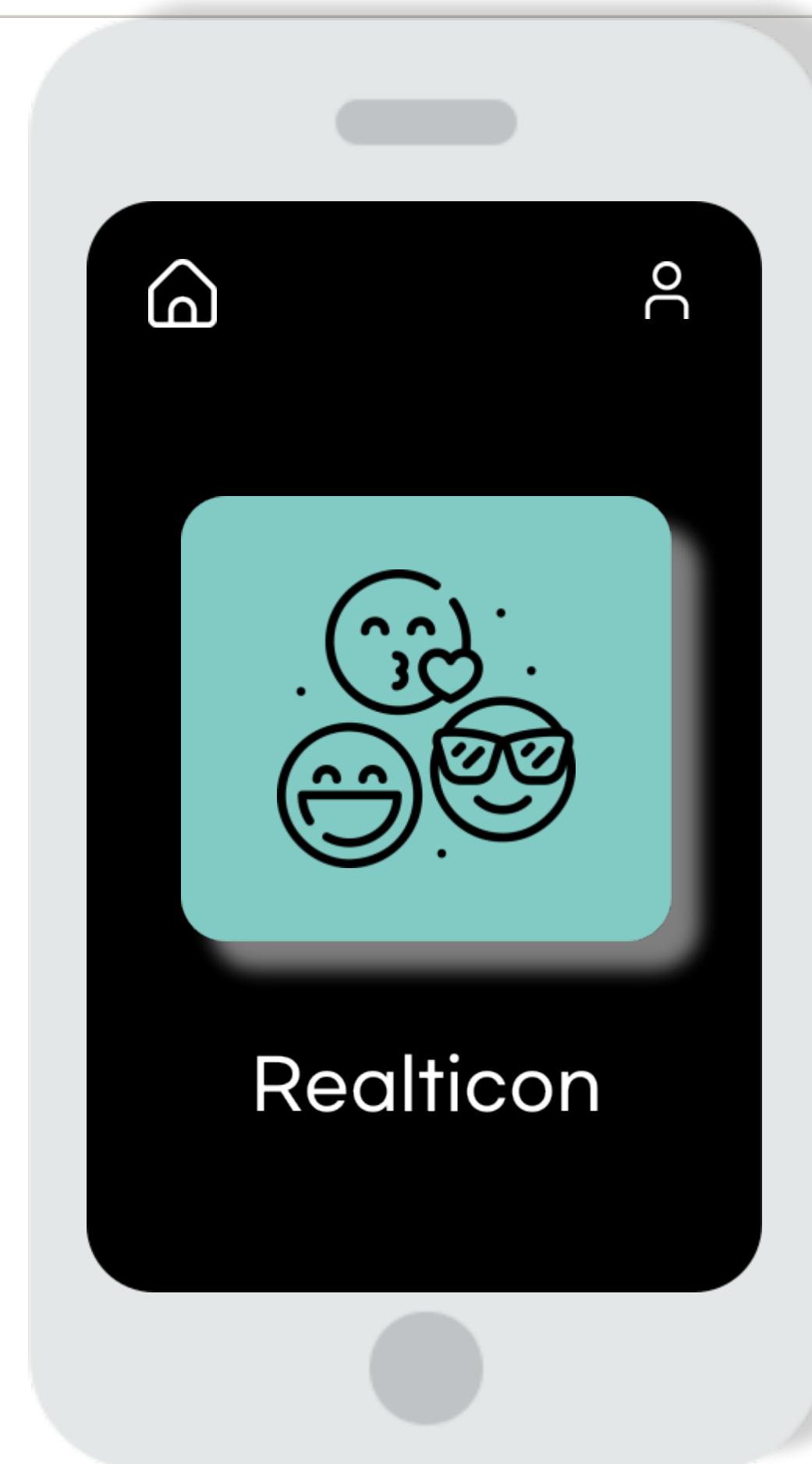
"밥잘챙겨먹고 건강해야해"와 같은 커스텀 영상메세지 서비스



driving.mp4

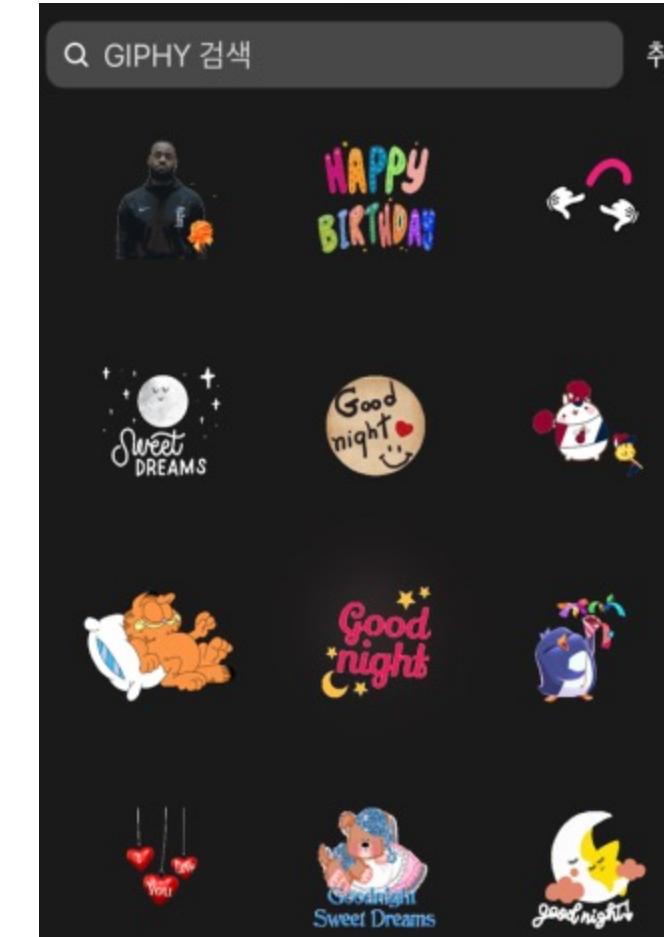
Audio.wav

Output



4. 리얼타운

리얼티콘 AI 이모티콘 생성 서비스



차별점

- 애플의 미모티콘처럼 내 아바타 만들기
- 일러스터, 연예인으로 만드는 이모티콘 짤



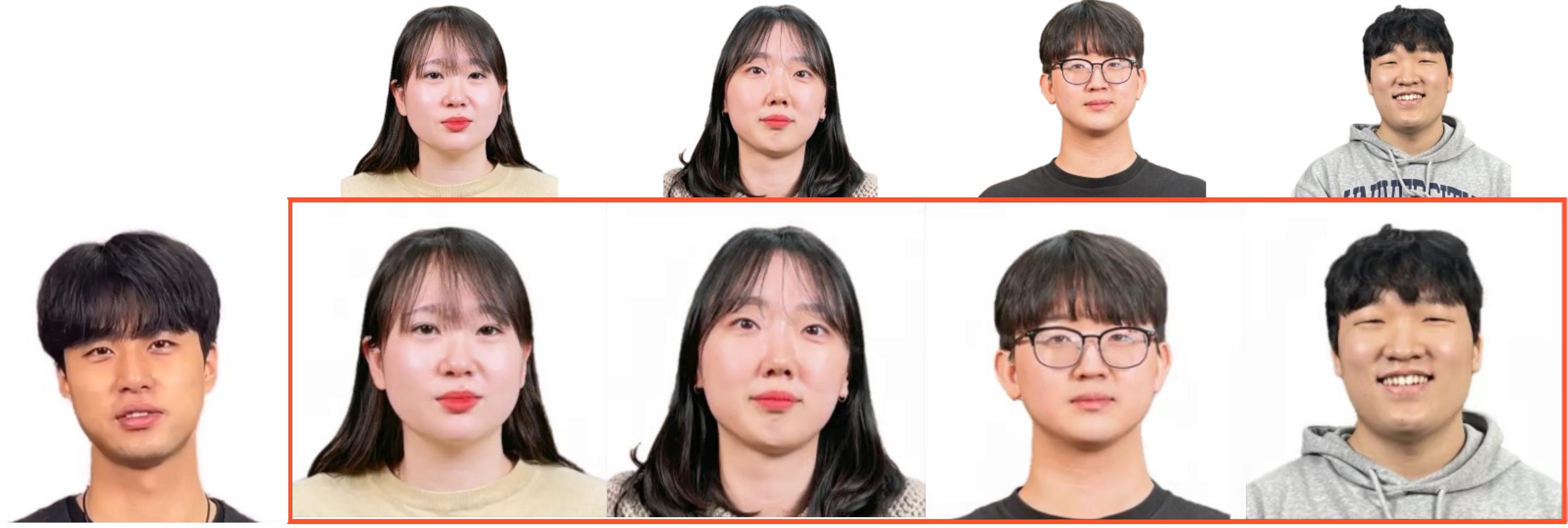
내 사진, 친구 사진, 연예인 사진 등 원하는 이미지로
원하는 표정, 말, 행동의 "이모티콘" 생성

리얼티콘

AI 이모티콘 생성 서비스

- AI가 사용자의 표정 및 신체 움직임을 학습하여 나만의 이모티콘을 생성

정신차려!



driving.mp4

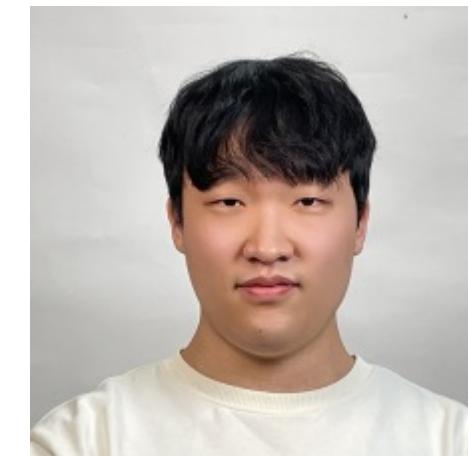
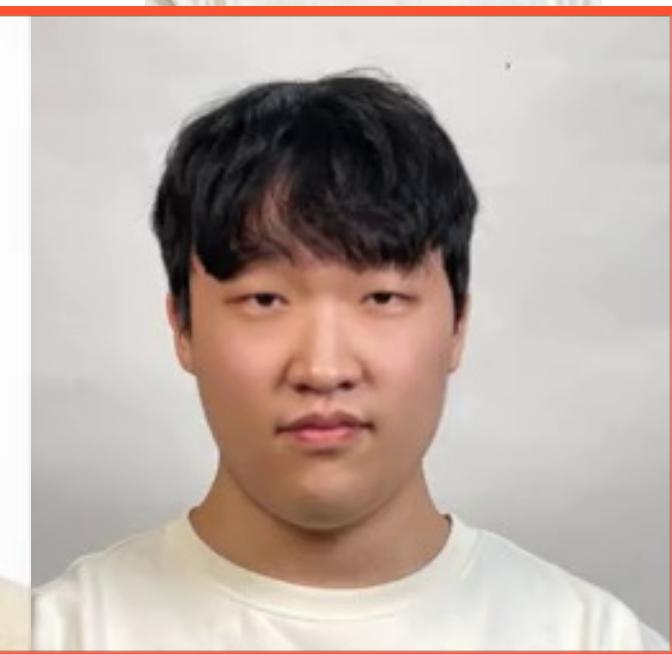
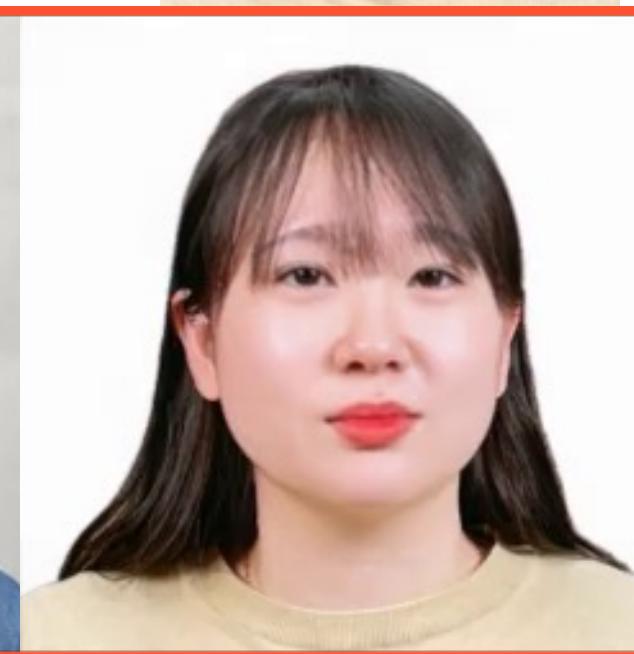
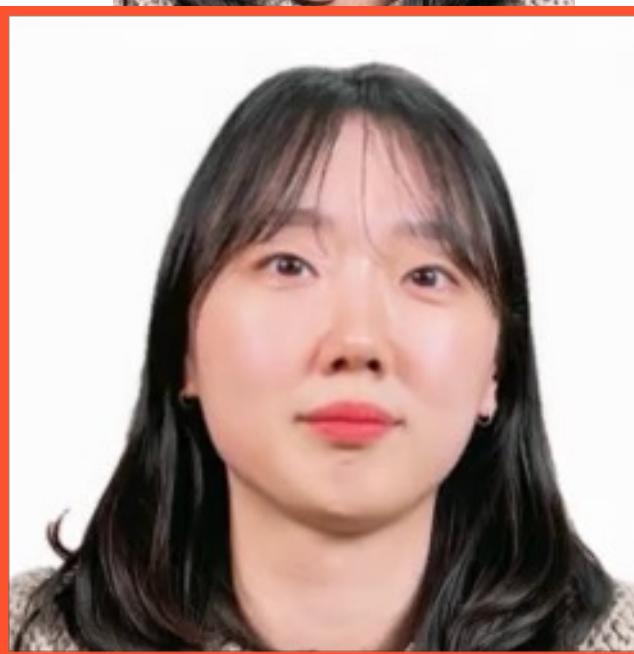
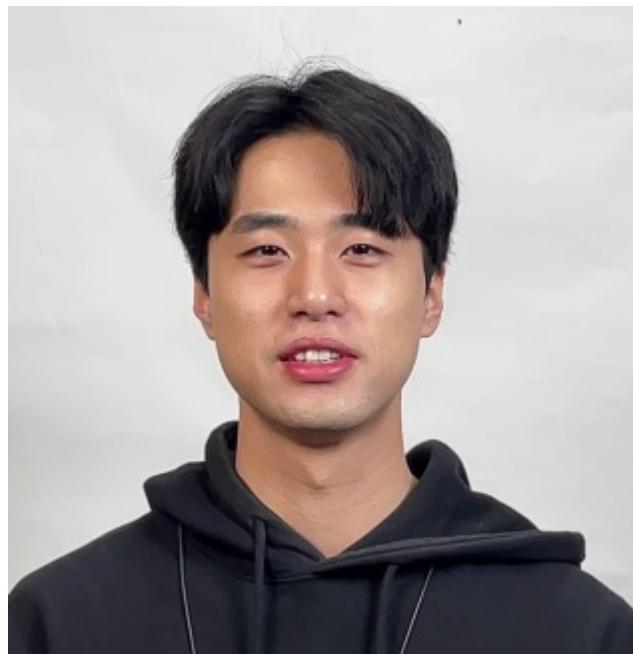
Output

리얼티콘

AI 이모티콘 생성 서비스

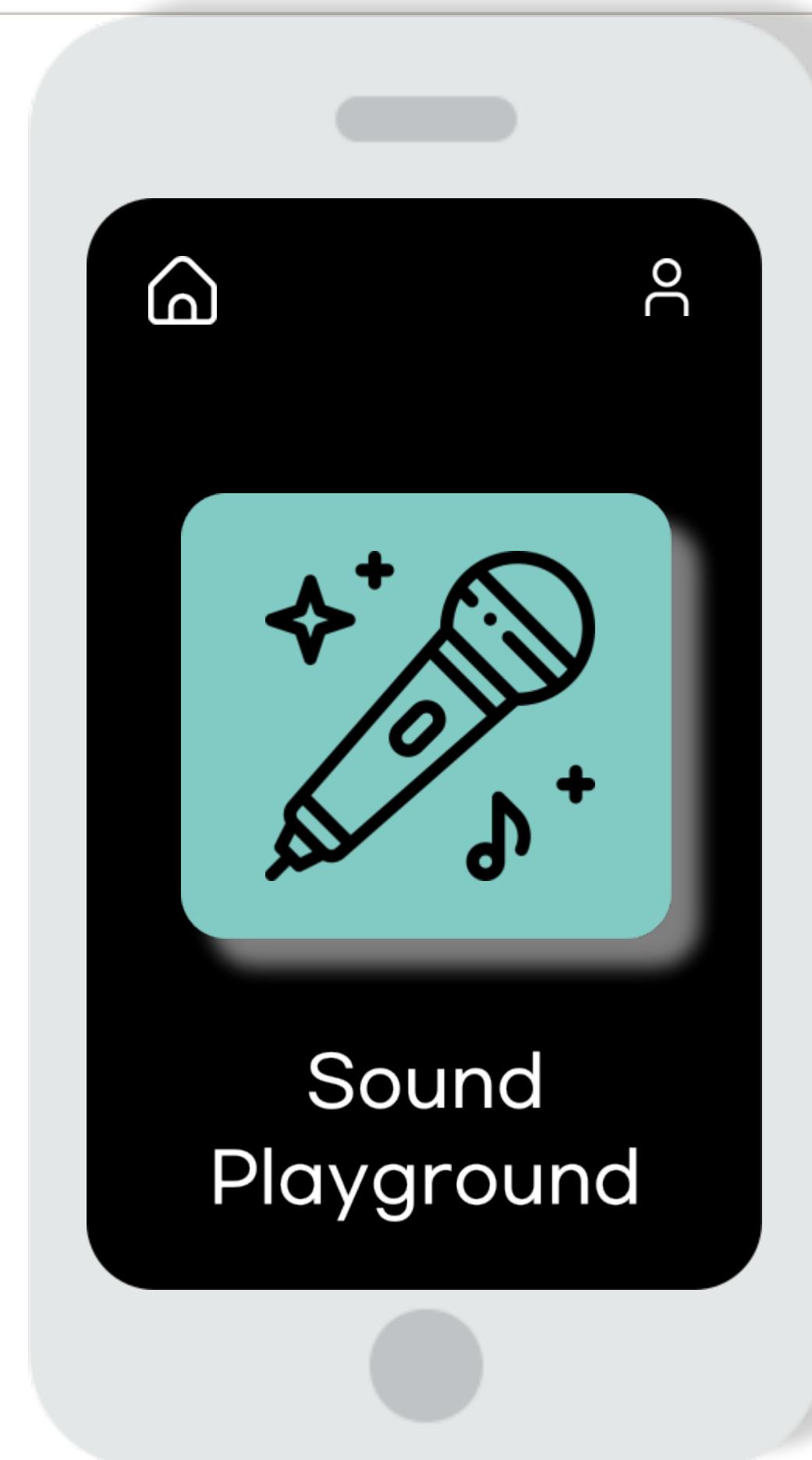
- AI가 사용자의 표정 및 신체 움직임을 학습하여 나만의 이모티콘을 생성

비상~ 초비상~!



driving.mp4

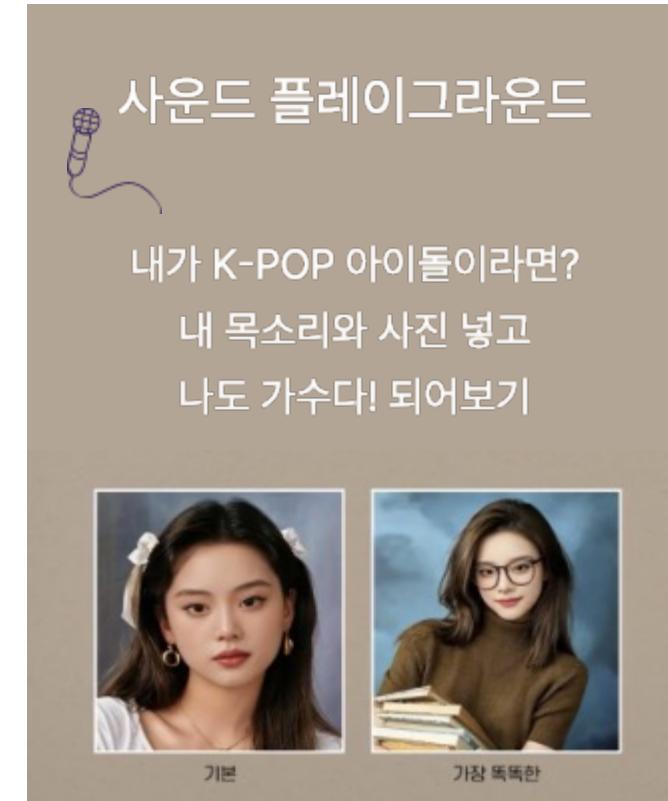
Output



5. 사운드 플레이어그라운드

사운드 플레이그라운드

나만의 AI 노래 생성 서비스



차별점

요즘 유행 중인 연예인의 AI 커버?
내 목소리로는 안되나?



내 목소리, 친구 목소리, 부모님 목소리, 연예인 목소리 등
원하는 목소리로 원하는 곡의 "노래 및 노래 영상" 생성

사운드 플레이어그라운드

나만의 AI 노래 생성 서비스

- 단 5분 길이의 음성으로 나도 가수가 될 수 있다!



Input



driving.mp4



▼ 10cm - 봄이좋냐



사운드 플레이어그라운드

나만의 AI 노래 생성 서비스

- 단 5분 길이의 음성으로 나도 가수가 될 수 있다!



Input



driving.mp4

▼

🎵 비비 - 밤양갱

♥

00:00 ~01:57

▷ ⏪ ⏩ ⏴



6. 가치 및 의의

- 가치 및 의의
- 발전방향 및 개선사항

가치 및 의의

1

한 장의 이미지

- 각도에 상관없이 한장의 사진만 있다면 일반적인 움직임 뿐만 아니라, 특정 상황에서의 움직임도 만들 수 있다.
- 사람, 동물 등의 움직이는 생물체 모두 가능
- 비지도학습

2

한국어로 가능

기존 TTS 모델은 영어 위주로, 한국어 모델이 없을 뿐더러, 발음이 자연스럽지 않다는 단점 존재
→ 약간의 어색함은 있지만 한국어 사용 TTS 모델 학습

3

효율적인 자원 활용

- 적은 소요 시간과 적은 비용으로 고객 개인 맞춤 서비스 제공
- 광고, 프리미엄, 방송, 영화 등 다수의 수익 창출 기회

개선사항

서비스 차원의 개선

- 서비스 기능 확대 : 오디오북, AI 챗봇 등
GPU 용량, 시간 → 모델 경량화
- 모바일로 서비스를 이용할 수 있도록 모델 개선
 - 지속적인 서비스 운영을 위한
수익 파이프라인 구축

기술적 차원의 개선

- TTS 성능 개선 : 한국어 발음 성능 개선
- 모델 경량화를 통해 on-device 적용
- 메모리 최적화 : GPU 용량 문제 해결 및
학습 시간 단축을 위한 개선/최적화 필요

발전방향



기존 모델 디벨롭 진행 중

>> 얼굴 뿐만 아니라 상반신으로 범위 확대

Reference

- 1. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech**
- 2. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold**
- 3. Motion Representations for Articulated Animation**
- 4. Thin-Plate Spline Motion Model for Image Animation**
- 5. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis**
- 6. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech**
- 7. CONTENTVEC: An Improved Self-Supervised Speech Representation by Disentangling Speakers**



Thank You!

경청해주세요 감사합니다.