

# 하이브리드 지능형 가상 캐릭터 동작 생성 개발

201724603 최우창

201724623 황원식

201724483 박현성

지도 교수: 이명호 교수님

# CONTENTS

1. 과제 개요
2. 과제 목표
3. 과제 내용
4. 결과 분석 및 평가
5. 결론 및 향후 연구 방향

# 과제 개요

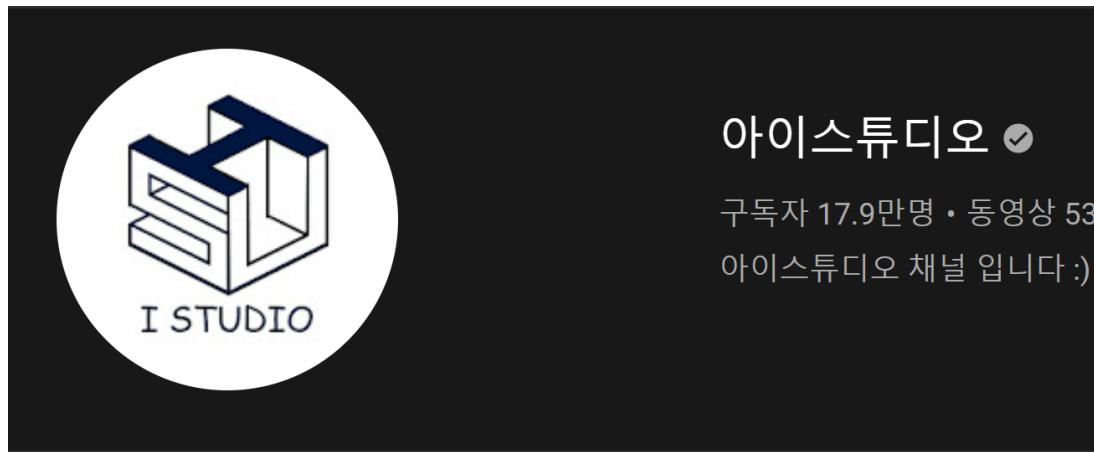
- 기존의 생성 모델 중 한글 텍스트 기반 동작 생성 모델이 존재하지 않았기에 한글 텍스트 데이터 셋 생성과 모델 구성을 수행한다.
  - 하이브리드 지능형 가상 캐릭터의 한글 텍스트 기반 동작 생성을 개발하는 것을 목표로 둔다.
-

# 과제 목표

- 한글 텍스트 데이터 셋 구축을 위해 유튜브 채널을 통해 영상과 자막을 얻는다.
  - OpenPose 라이브러리를 이용하여 2차원 Pose 추출 후 3차원 Pose 추정 작업을 진행한다.
  - 해당 데이터 셋을 모델에 학습시켜 한글 텍스트에 대한 제스처를 생성한다.
-

# 과제 내용 - 학습 데이터 셋 생성

## 1. 유튜브 영상 및 자막 웹 크롤링



영상 및 자막을 웹 크롤링한 채널  
<아이 스튜디오>

```
00:00:31.999 --> 00:00:32.009 align:start position:0%
바라보고 있는 겁니다 미래의 내가

00:00:32.009 --> 00:00:34.700 align:start position:0%
바라보고 있는 겁니다 미래의 내가
어떤 <00:00:32.296><c>일</c><00:00:32.583><c>이 </c><00:00:32.870><c>벌어질까 </c>

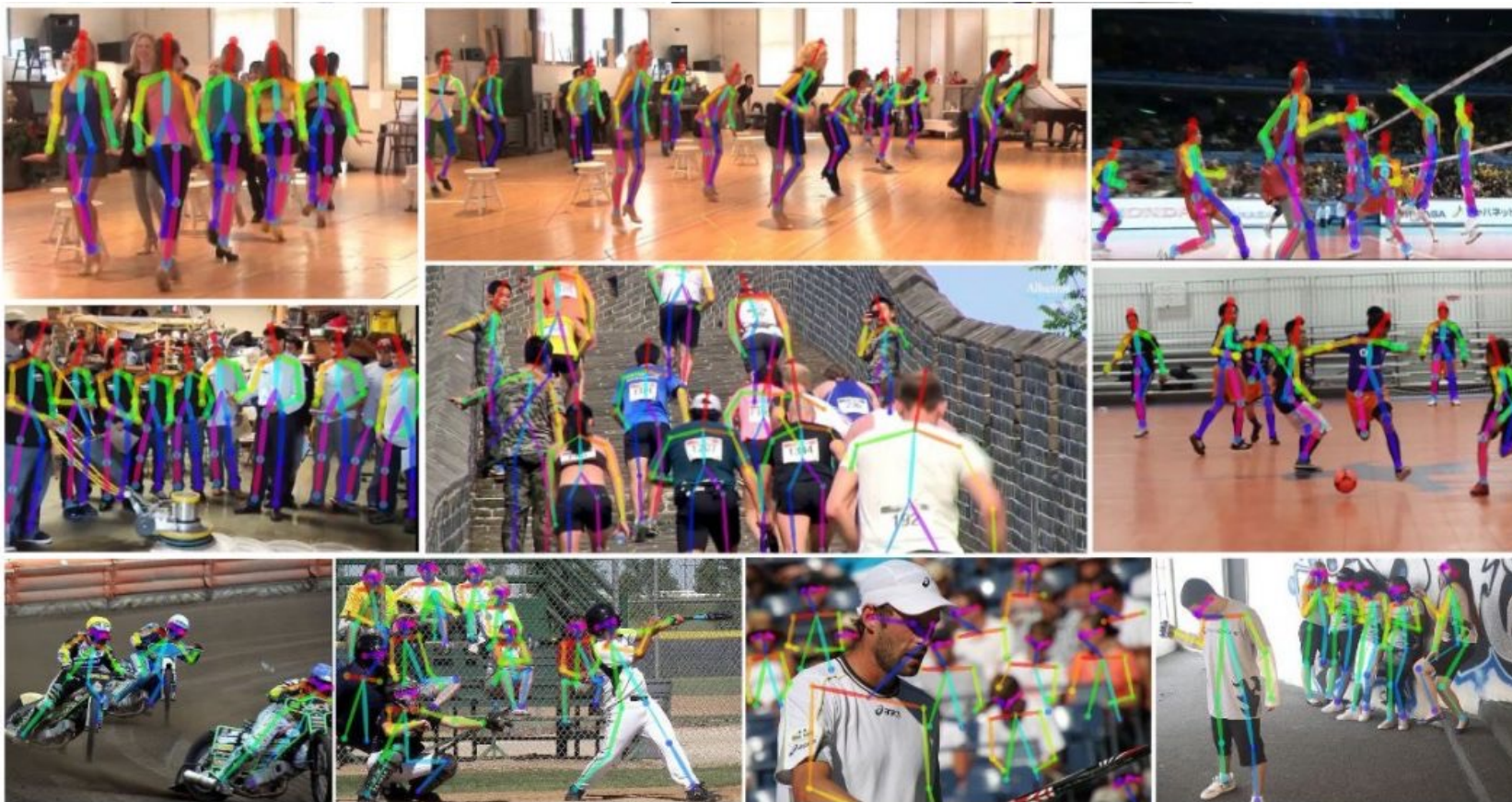
00:00:34.700 --> 00:00:34.710 align:start position:0%
어떤 일이 벌어질까 봐 그걸 준비해야

00:00:34.710 --> 00:00:36.770 align:start position:0%
어떤 일이 벌어질까 봐 그걸 준비해야
돼 <00:00:34.995><c>아니면 </c><00:00:35.280><c>그 </c><00:00:35.565><c>준비</c>
```

추출한 자막 파일

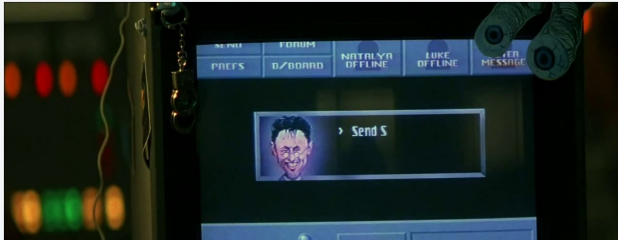

# 과제 내용 - 학습 데이터 셋 생성

## 2. OpenPose를 이용해 2D Human Pose 추출



# 과제 내용 - 학습 데이터 셋 생성

## 3. PySceneDetect로 영상을 클립으로 나눈 뒤 유효한 클립 선별

Scene #	Start Time	Preview
1	00:00:00.000	
3	00:00:08.759	

영상 클립 분할

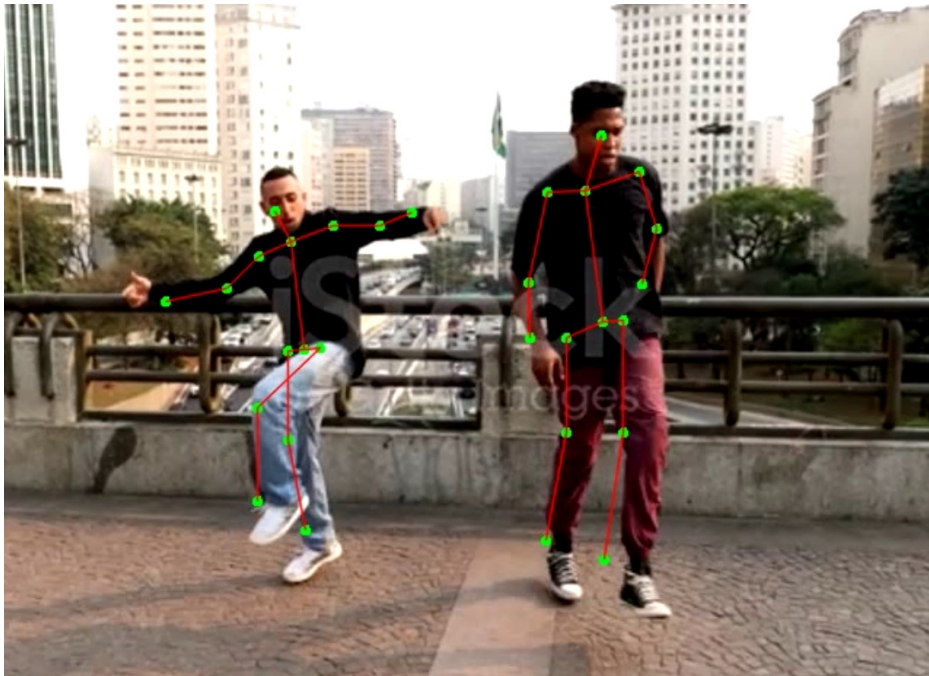
```
[{"clip_info": [0, 69, false], "filtering_results": [0, 0, 0, 0, 0, 0, 0], "message": "too Short", "debugging_info": ["None", "None", "None", "None", "None"]}, {"clip_info": [69, 149, true], "filtering_results": [1, 1, 1, 1, 1, 1, 1], "message": "PASS", "debugging_info": [0.0, 0.0, 0.0, 0.0, 6407063.0]}, {"clip_info": [149, 237, false], "filtering_results": [1, 1, 1, 0, 0, 0, 0], "message": "too many missing joints", "debugging_info": [0.0, 1.0, "None", "None", "None"]}, {"clip_info": [237, 293, false], "filtering_results": [0, 0, 0, 0, 0, 0, 0], "message": "too Short", "debugging_info": ["None", "None", "None", "None", "None"]}, {"clip_info": [293, 376, false], "filtering_results": [1, 1, 1, 0, 0, 0, 0], "message": "too many missing joints", "debugging_info": [0.0, 1.0, "None", "None", "None"]}, {"clip_info": [376, 620, true], "filtering_results": [1, 1, 1, 1, 1, 1, 1], "message": "PASS", "debugging_info": [0.0, 0.373, 0.0, 0.0, 54920435.0]}, {"clip_info": [620, 756, true], "filtering_results": [1, 1, 1, 1, 1, 1, 1], "message": "PASS", "debugging_info": [0.103, 0.103, 0.103, 0.103, 10963778.0]}, {"clip_info": [756, 918, false], "filtering_results": [1, 1, 1, 0, 0, 0, 0], "message": "too many missing joints", "debugging_info": [0.0, 0.802, "None", "None", "None"]}, {"clip_info": [918, 1267, true], "filtering_results": [1, 1, 1, 1, 1, 1, 1], "message": "PASS", "debugging_info": [0.0, 0.0, 0.0, 0.0, 42732100.0]}, {"clip_info": [1267, 1374, false], "filtering_results": [0, 0, 0, 0, 0, 0, 0], "message": "too Short", "debugging_info": ["None", "None", "None", "None", "None"]}]
```

유효한 클립 정보

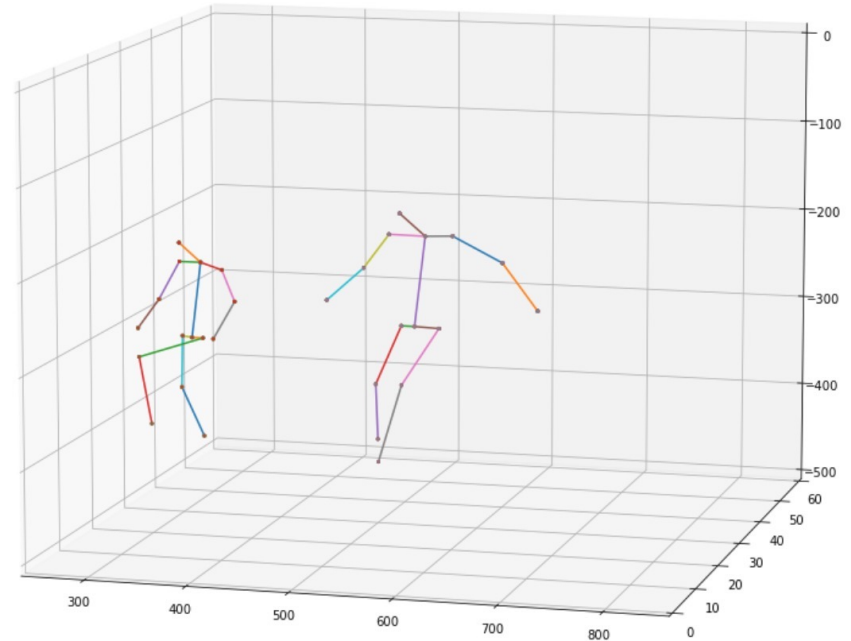


# 과제 내용 - 학습 데이터 셋 생성

## 4. 유효한 클립의 2D Pose를 3D Pose로 추정



기존 2D Pose



추정된 3D Pose



# 과제 내용 - 학습 데이터 셋 생성

## 5. 추출된 데이터를 통해 LMDB 데이터 셋 생성

```
dtype=float32), 'zuYz0n0U2PY_clip068': array([[[-3.7438624e-06, -2.2638276e-05, 3.6349040e-05],  
 [ 5.6861932e-03, -2.5184566e-01, 4.4073448e-02],  
 [ 4.0685140e-02, -3.5330948e-01, 9.6983343e-02],  
 ...,  
 [ 1.3203235e-01, -1.9006559e-01, 1.2857139e-02],  
 [ 2.4576584e-01, 6.5480910e-02, 4.7942575e-02],  
 [ 2.7066320e-01, -6.7278549e-02, 2.3875567e-01]],  
  
 [[-3.7970601e-06, -2.2554193e-05, 3.5995548e-05],  
 [ 5.4821419e-03, -2.5199121e-01, 4.4124883e-02],  
 [ 3.9480854e-02, -3.5339662e-01, 9.6299686e-02],  
 ...,  
 [ 1.3189553e-01, -1.8960872e-01, 1.3939083e-02],  
 [ 2.4343815e-01, 6.5705433e-02, 4.9095817e-02],  
 [ 2.7271396e-01, -6.6182308e-02, 2.3972911e-01]],  
  
 [[-3.8261887e-06, -2.2845858e-05, 3.6422174e-05],  
 [ 5.1991604e-03, -2.5242496e-01, 4.4236537e-02],  
 [ 3.8042013e-02, -3.5363951e-01, 9.5158175e-02],  
 ...,  
 [ 1.3131969e-01, -1.8842302e-01, 1.6339041e-02],  
 [ 2.4056999e-01, 6.6475078e-02, 5.0906900e-02],  
 [ 2.7558777e-01, -6.4377889e-02, 2.4174240e-01]]],
```

LMDB 데이터 셋 내부 형태

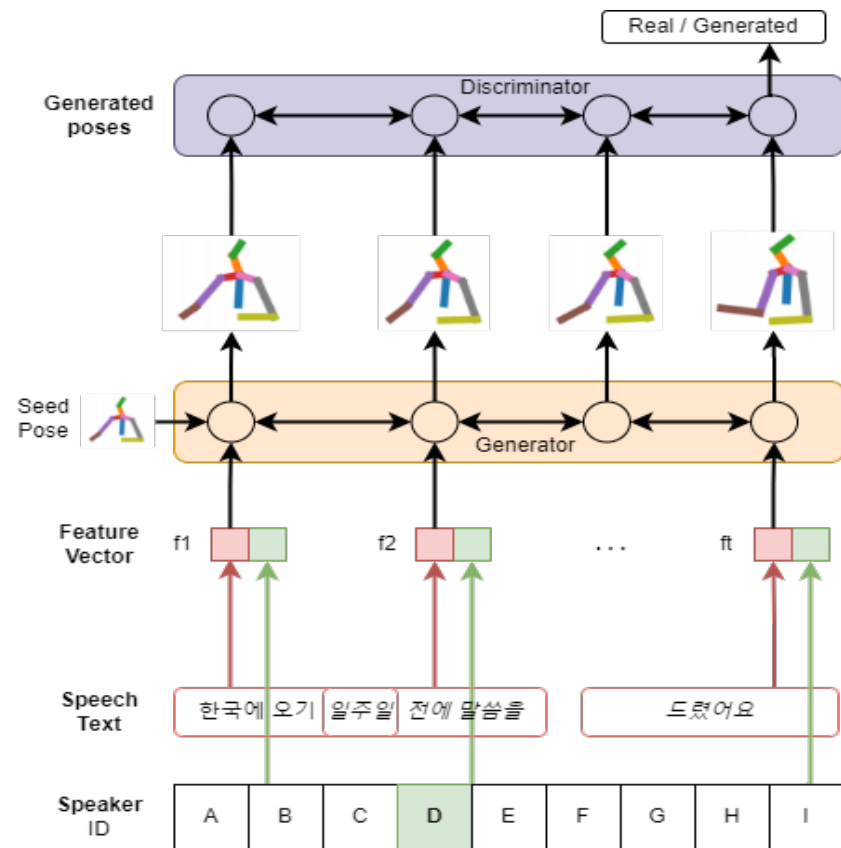
사용한 비디오 수	328개
비디오 평균 길이	6.4분
전체 비디오 프레임 수	3857343 frames
사용한 프레임 비율	28%(1094727/3857343)
사용 비디오 길이	10.03시간

사용한 데이터 양

# 과제 내용 - 제스처 생성 모델 학습 및 구조

```
■ Anaconda Prompt (anaconda3) - conda deactivate - python scripts/train.py --config=config/multimodal_context.yml
2022-09-21 17:03:10.403: EP 31 (220) | 325m 41s, 40 samples/s | loss: 19.191, gen: 3.952, dis: 1.319, KLD: 0.145, DIV_REG: -2.536,
2022-09-21 17:03:34.617: [VAL] loss: 0.086, joint mae: 0.03143, accel diff: 0.00033, FGD: 46.277, feat_D: 128.373 / 11.9s
2022-09-21 17:03:34.618: best validation loss so far: 30.508 at EPOCH 31
2022-09-21 17:05:52.057: EP 32 ( 44) | 328m 23s, 41 samples/s | loss: 18.255, gen: 3.941, dis: 1.326, KLD: 0.145, DIV_REG: -2.548,
2022-09-21 17:06:09.421: EP 32 ( 88) | 330m 40s, 41 samples/s | loss: 18.570, gen: 3.976, dis: 1.318, KLD: 0.145, DIV_REG: -2.562,
2022-09-21 17:10:26.767: EP 32 (132) | 332m 57s, 41 samples/s | loss: 18.552, gen: 3.946, dis: 1.323, KLD: 0.146, DIV_REG: -2.523,
2022-09-21 17:12:44.022: EP 32 (176) | 335m 15s, 41 samples/s | loss: 18.820, gen: 3.934, dis: 1.324, KLD: 0.146, DIV_REG: -2.548,
2022-09-21 17:15:01.472: EP 32 (220) | 337m 32s, 40 samples/s | loss: 18.744, gen: 3.919, dis: 1.320, KLD: 0.146, DIV_REG: -2.642,
2022-09-21 17:15:25.978: [VAL] loss: 0.090, joint mae: 0.03252, accel diff: 0.01041, FGD: 35.153, feat_D: 132.568 / 11.8s
2022-09-21 17:15:25.980: best validation loss so far: 30.508 at EPOCH 31
2022-09-21 17:17:43.328: EP 33 ( 44) | 340m 14s, 42 samples/s | loss: 18.135, gen: 3.940, dis: 1.317, KLD: 0.146, DIV_REG: -2.575,
2022-09-21 17:20:01.174: EP 33 ( 88) | 342m 32s, 43 samples/s | loss: 18.163, gen: 3.967, dis: 1.325, KLD: 0.147, DIV_REG: -2.556,
2022-09-21 17:22:18.494: EP 33 (132) | 344m 49s, 43 samples/s | loss: 18.553, gen: 3.978, dis: 1.321, KLD: 0.147, DIV_REG: -2.567,
2022-09-21 17:24:35.750: EP 33 (176) | 347m 6s, 42 samples/s | loss: 18.510, gen: 3.968, dis: 1.318, KLD: 0.147, DIV_REG: -2.601,
2022-09-21 17:26:53.483: EP 33 (220) | 349m 24s, 41 samples/s | loss: 18.461, gen: 3.941, dis: 1.314, KLD: 0.147, DIV_REG: -2.603,
2022-09-21 17:27:17.648: [VAL] loss: 0.089, joint mae: 0.03228, accel diff: 0.00859, FGD: 35.661, feat_D: 132.264 / 11.9s
2022-09-21 17:27:17.649: best validation loss so far: 30.508 at EPOCH 31
2022-09-21 17:29:34.891: EP 34 ( 44) | 352m 6s, 41 samples/s | loss: 17.889, gen: 3.889, dis: 1.330, KLD: 0.148, DIV_REG: -2.637,
2022-09-21 17:31:52.480: EP 34 ( 88) | 354m 23s, 41 samples/s | loss: 18.139, gen: 3.963, dis: 1.315, KLD: 0.148, DIV_REG: -2.618,
2022-09-21 17:34:10.223: EP 34 (132) | 356m 41s, 41 samples/s | loss: 18.276, gen: 3.948, dis: 1.322, KLD: 0.148, DIV_REG: -2.638,
2022-09-21 17:36:27.594: EP 34 (176) | 358m 58s, 41 samples/s | loss: 18.382, gen: 3.947, dis: 1.320, KLD: 0.148, DIV_REG: -2.658,
2022-09-21 17:38:45.208: EP 34 (220) | 361m 16s, 41 samples/s | loss: 18.237, gen: 3.903, dis: 1.323, KLD: 0.150, DIV_REG: -2.644,
2022-09-21 17:39:09.654: [VAL] loss: 0.090, joint mae: 0.03264, accel diff: 0.00981, FGD: 23.476, feat_D: 136.065 / 11.8s
2022-09-21 17:39:09.655: *** BEST VALIDATION LOSS: 23.476
2022-09-21 17:39:09.654: Saved the checkpoint
2022-09-21 17:41:27.444: EP 35 ( 44) | 363m 58s, 41 samples/s | loss: 17.855, gen: 3.913, dis: 1.326, KLD: 0.149, DIV_REG: -2.650,
2022-09-21 17:43:46.702: EP 35 ( 88) | 366m 17s, 39 samples/s | loss: 18.017, gen: 3.962, dis: 1.321, KLD: 0.150, DIV_REG: -2.689,
2022-09-21 17:46:05.318: EP 35 (132) | 368m 36s, 41 samples/s | loss: 18.048, gen: 3.929, dis: 1.325, KLD: 0.150, DIV_REG: -2.685,
2022-09-21 17:48:23.733: EP 35 (176) | 370m 54s, 40 samples/s | loss: 18.257, gen: 3.897, dis: 1.325, KLD: 0.151, DIV_REG: -2.710,
```

Model 학습 진행



설계한 Model 구조

# 과제 내용 - 제스처 생성 모델 학습 및 구조

$$L_G = \alpha \cdot L_G^{\text{Huber}} + \beta \cdot L_G^{\text{NSGAN}} + \gamma \cdot L_G^{\text{style}} + \lambda \cdot L_G^{\text{KLD}} \quad (1)$$

$$L_G^{\text{Huber}} = \mathbb{E}\left[\frac{1}{t} \sum_{i=1}^t \text{HuberLoss}(d_i, \hat{d}_i)\right] \quad (2)$$

$$L_G^{\text{NSGAN}} = -\mathbb{E}[\log(D(\hat{d}))] \quad (3)$$

$$L_G^{\text{style}} = -\mathbb{E}\left[\min\left(\frac{\text{HuberLoss}(G(f^{\text{text}}, f^{\text{style}_1}) - G(f^{\text{text}}, f^{\text{style}_2}))}{\|f^{\text{style}_1} - f^{\text{style}_2}\|_1}, \tau\right)\right] \quad (4)$$

$$L_D = -\mathbb{E}[\log(D(d))] - \mathbb{E}[\log(1 - D(\hat{d}))] \quad (5)$$

$d$ : Pose

$\hat{d}$ : Generated Pose

$L_G$ : Gesture Generator와 Encoder를 학습시킬 때 사용되는 값

$L_D$ : Discriminator를 학습시킬 때 사용되는 값

$L_G^{\text{NSGAN}}, L_D$ : 서로 Adversarial Loss이다

$L_G^{\text{style}}$ : Style Feature

$L_G^{\text{KLD}}$ : (0,1) 정규분포와 정규분포로 추정된 style embedding space 사이의 차이를 바탕으로 style embedding space의 분포가 너무 퍼지지 않게끔 하는 역할

# 과제 내용 - 제스처 생성 모델 학습 및 구조

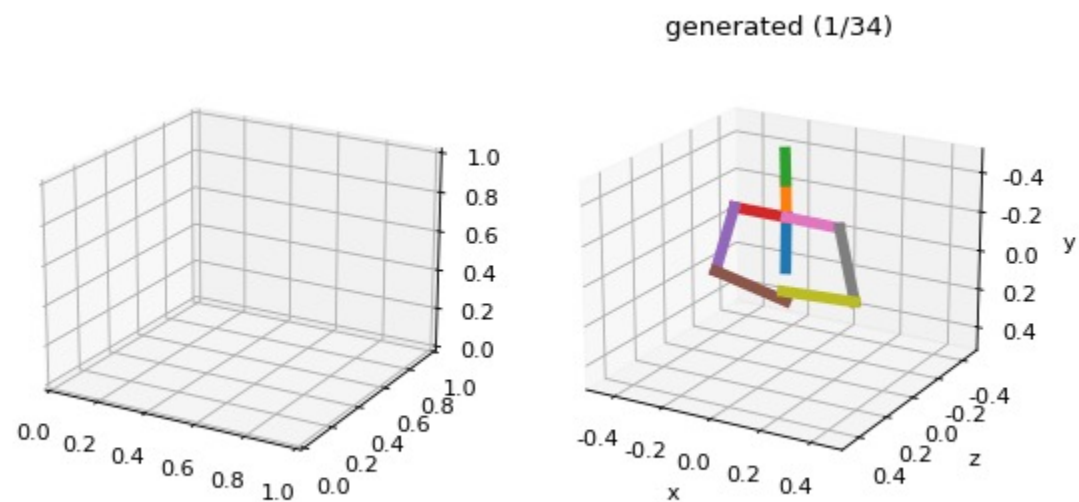
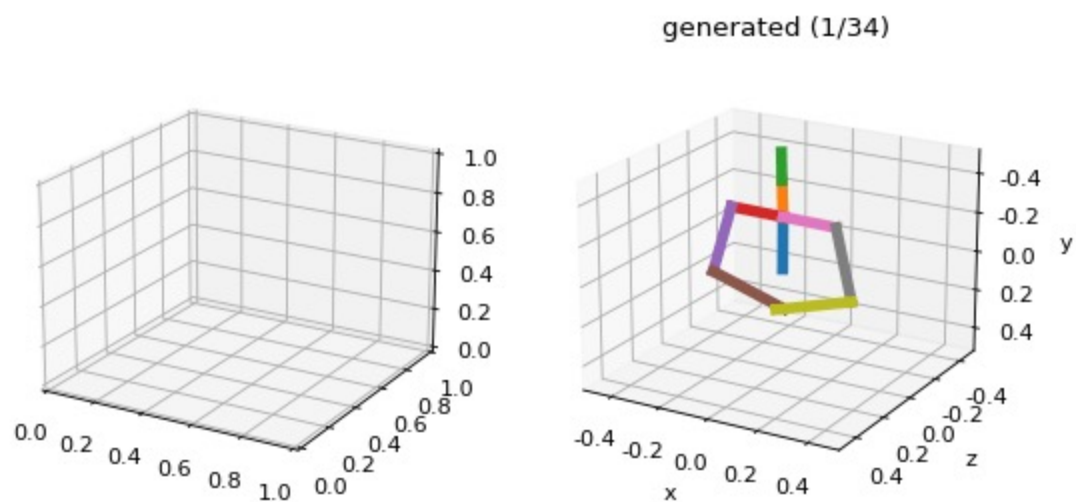
[['안녕하세요', 0.0, 0.6], ['오늘은', 0.6, 1.5], ['셰프가', 1.5, 2.0], ['준비한', 2.0, 2.3], ['요리를', 2.3, 2.8], ['대  
접해', 2.8, 3.2], ['드릴게요', 3.2, 3.4], ['맛있는', 3.4, 4.3], ['음식', 4.3, 4.9], ['기대해주세요', 4.9, 5.2]]

Google STT를 이용해 생성된 타임스탬프

# TTS



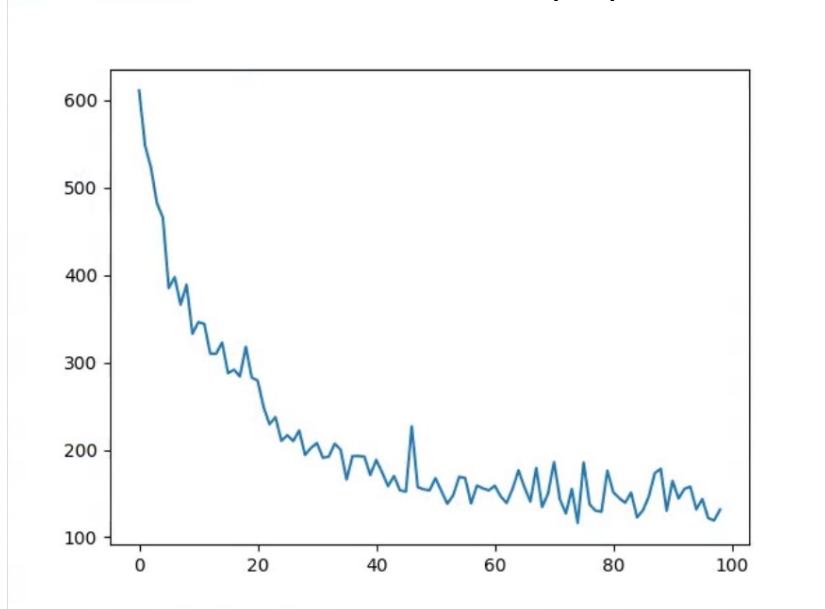
# 결과 분석 및 평가



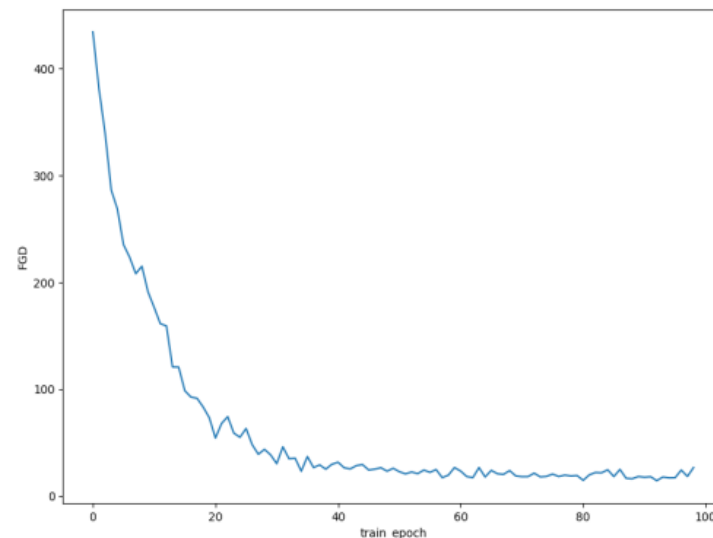
# 결과 분석 및 평가

- FGD

GAN 모델을 평가할 때 자주 사용되는 FID(Frechet Inception Distance)라는 생성된 이미지 분포와 기존 이미지 분포가 얼마나 유사한지 측정하는 지표를 Gesture Generation Problem에 적용한 값



**Sequence2Sequence**  
(Min FGD: 116.269)



**Proposed Model**  
(Min FGD: 14.637)

---

# 결론 및 향후 연구 방향



Clova<sup>ⓧ</sup>

