Human-Centric Visual Generation and Editing

Ziwei Liu 刘子纬

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Creative Industry

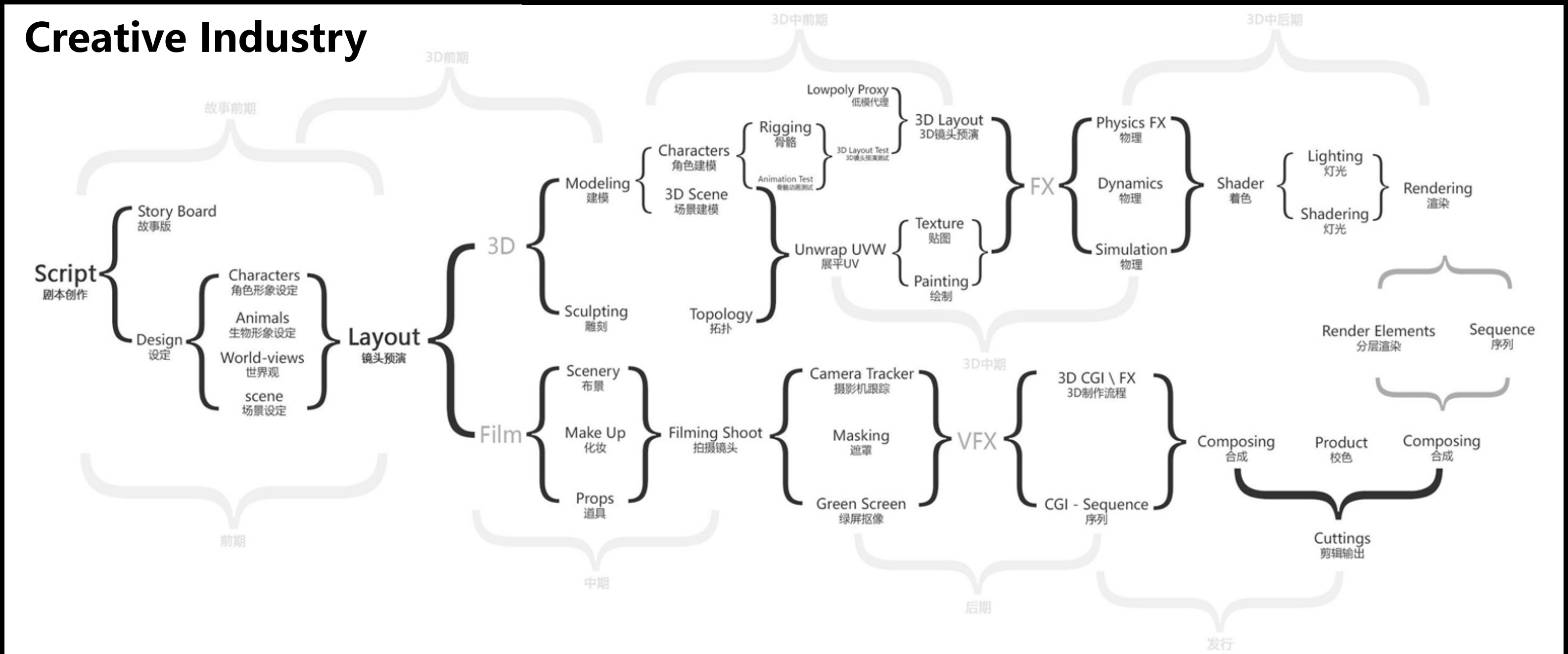
Movie

Game

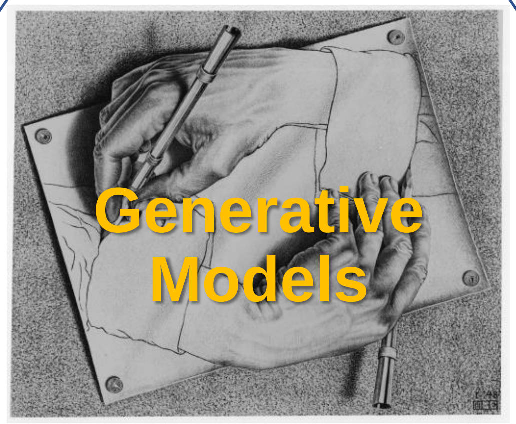
Anime

VTuber

Virtual Beings



Human-Centric



Scaling

Interactive

Scaling Generative Models

CelebV-HQ:
A Large-Scale Video Facial Attributes Dataset

Hao Zhu[1*], Wayne Wu[1*] , Wentao Zhu[2], Liming Jiang[3],
Siwei Tang[1], Li Zhang[1], Ziwei Liu[3], Chen Change Loy[3]
(Equal contribution)

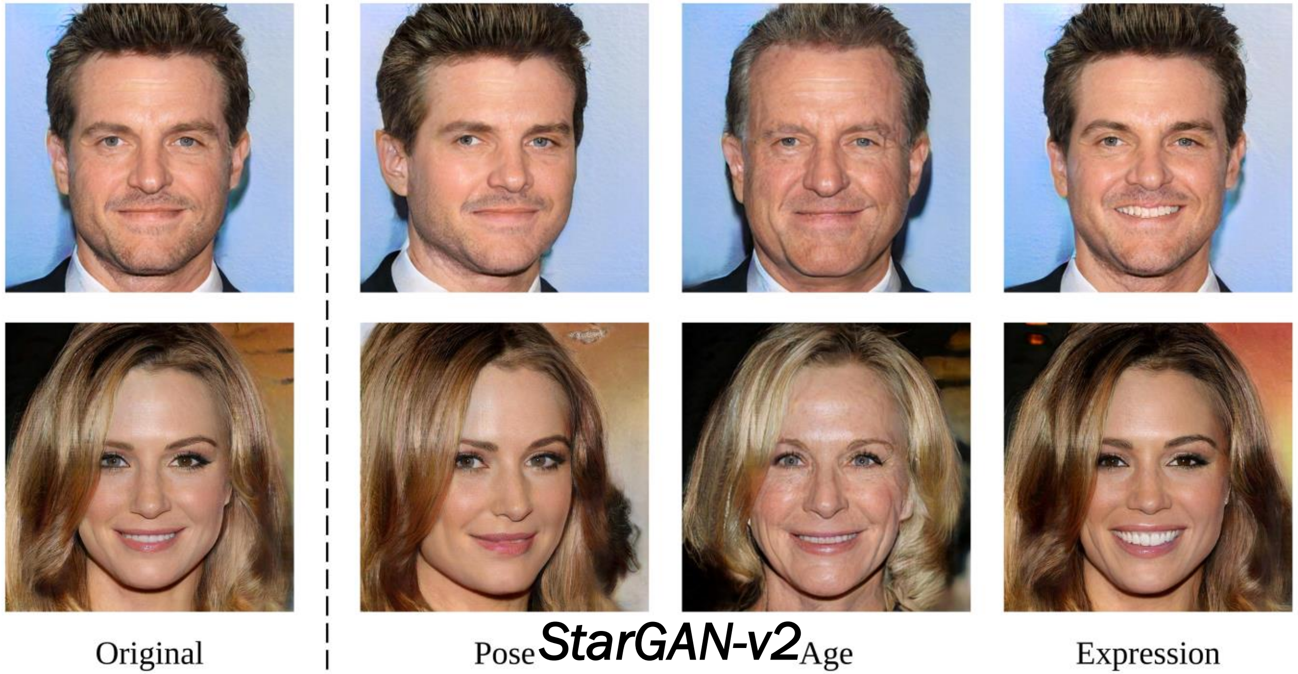
[1]SenseTime Research
[2]Peking University
[3]S-Lab, Nanyang Technological University

ECCV 2022

Motivation

- Large-scale datasets play an indispensable role in the recent successes of face generation and editing.
- The **practical applications** of powerful GANs have also been expanded in both academia and industry.

CelebA-HQ



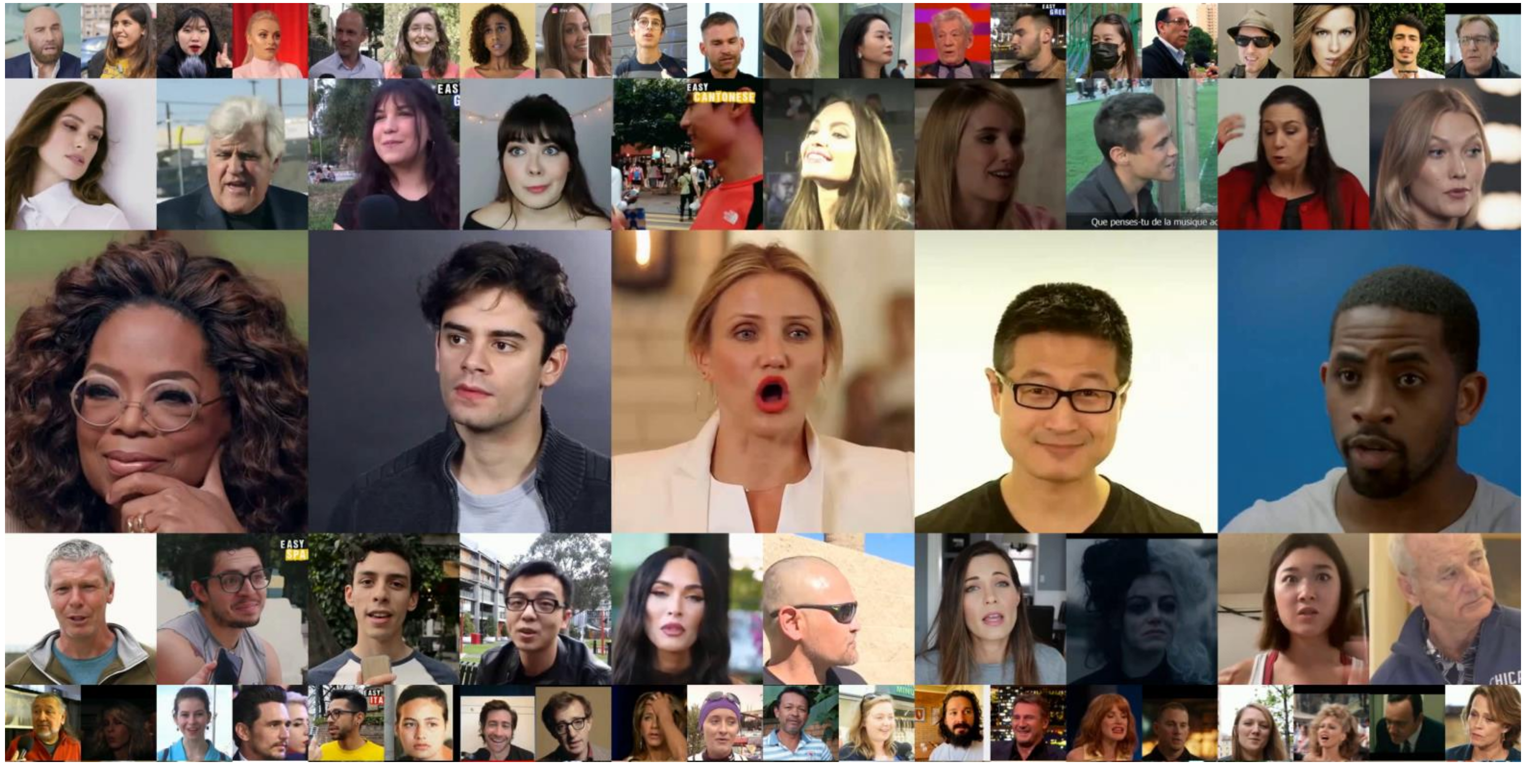
StarGAN-v2

Motivation

- A large-scale face video dataset with facial attributes is still missing...

CelebA-HQ

CelebV-HQ



- 35,666 video clips
- 15,653 IDs
- 83 attributes
  - 40 Appearance
  - 35 Action
  - 8 Emotion

Statistics of **CelebV-HQ**

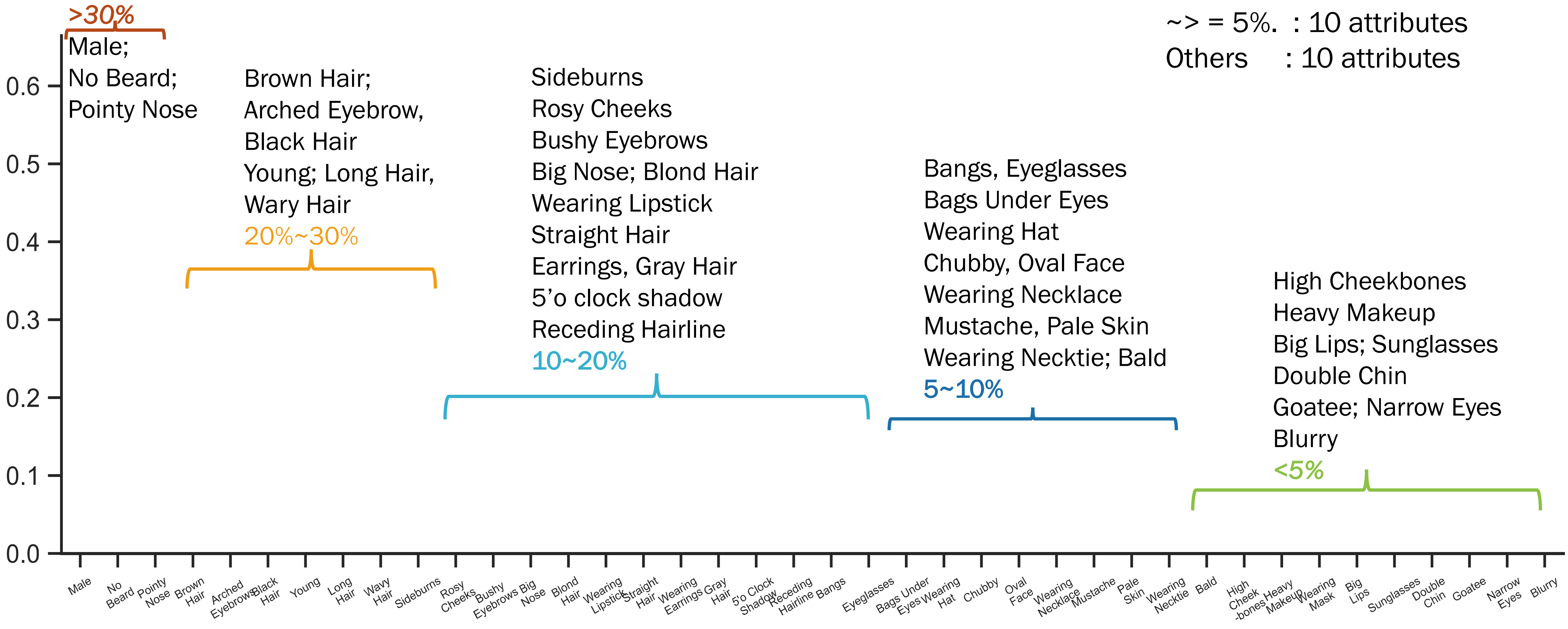
Appearance/Action/Emotion

CelebV-HQ: Analysis

Appearance

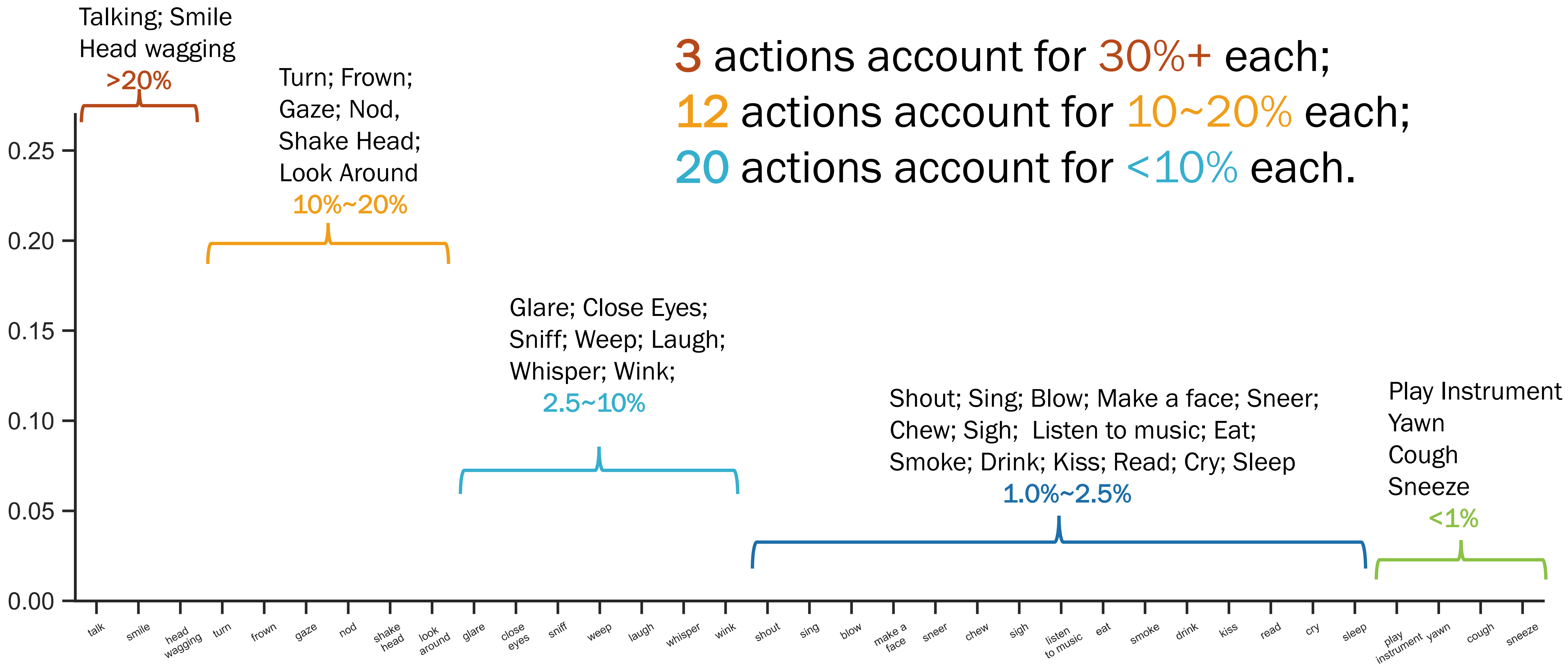
40 appearance attributes

~> = 20% : 10 attributes
~> = 10% : 10 attributes
~> = 5%. : 10 attributes
Others   : 10 attributes



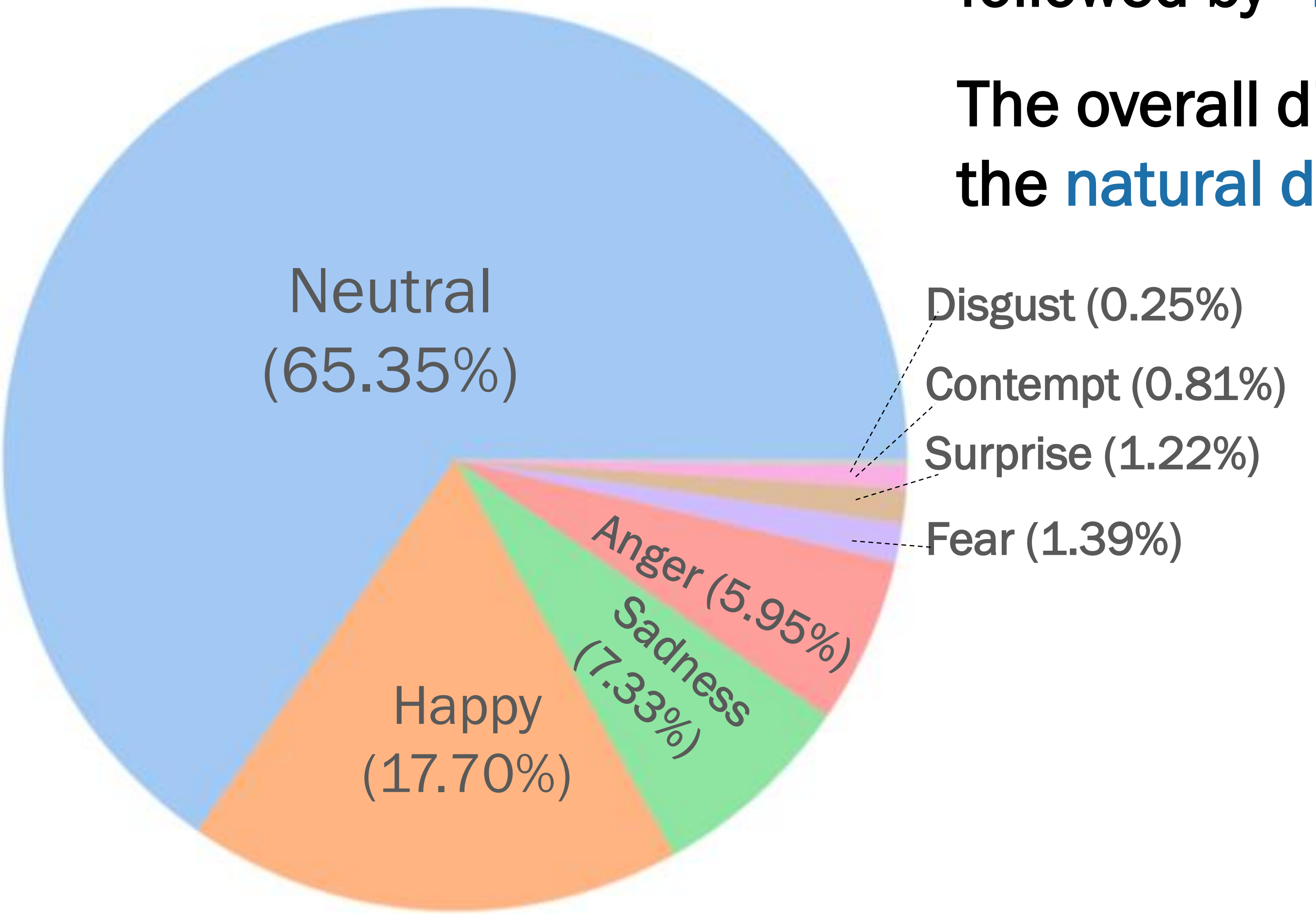
CelebV-HQ: Analysis

Action



CelebV-HQ: Analysis

Emotion



"Neutral" accounting for 65.35%, followed by "happiness" and "sadness"

The overall distribution is in line with the natural distribution.

Comparison with VoxCeleb

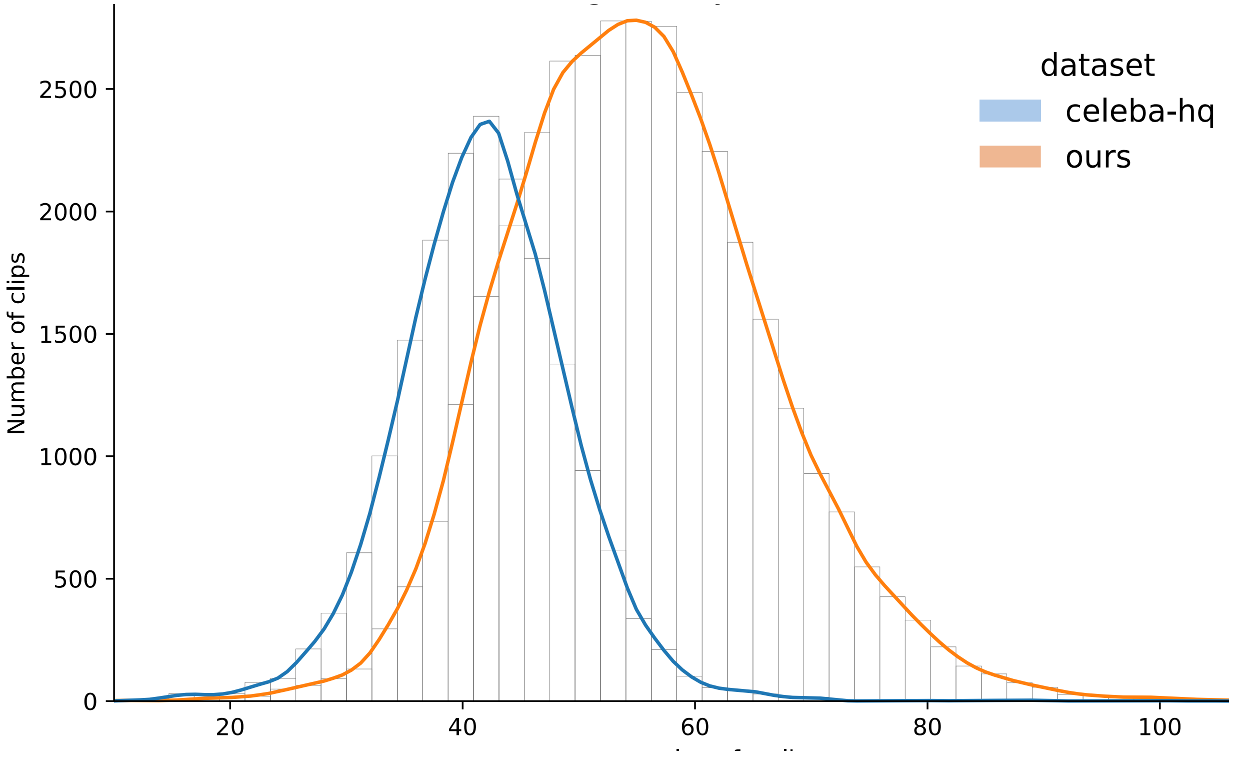
Quality/ Head Pose / Action Unit

Distribution Comparison

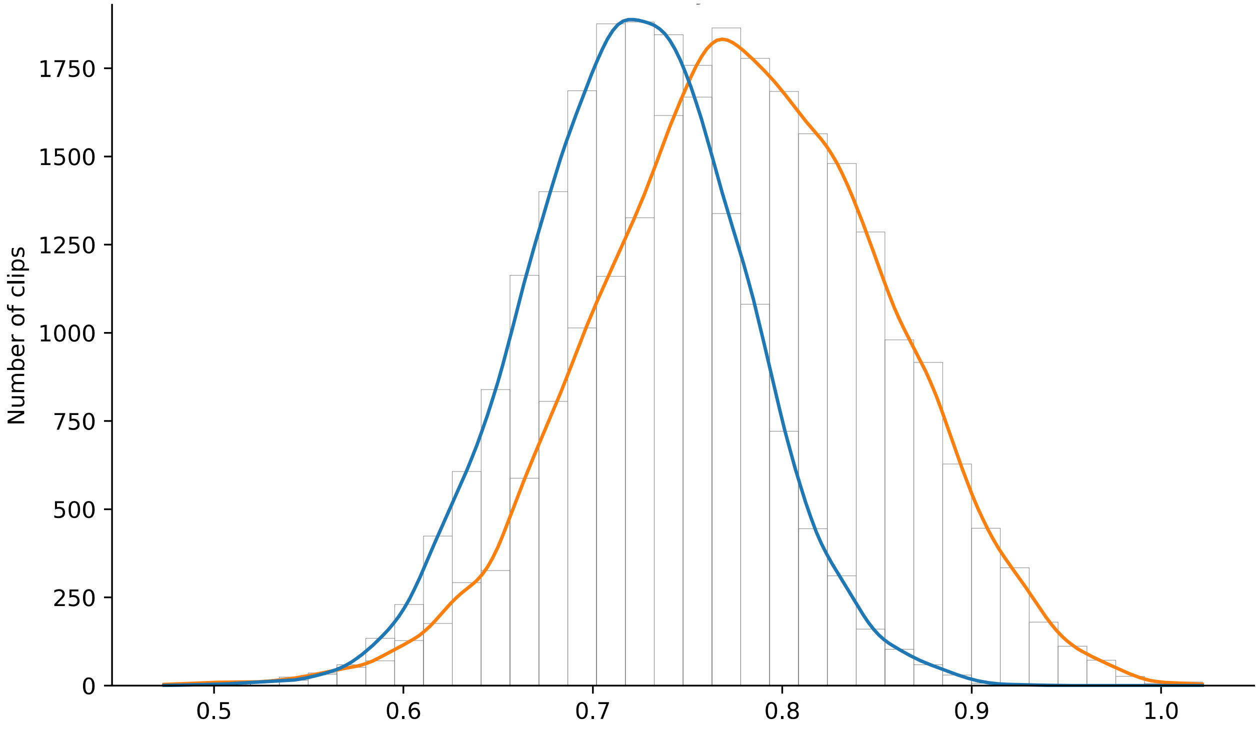
VoxCeleb – Image/Video Quality

CelebV-HQ achieves better performance

Image quality is measured by BRISQUE



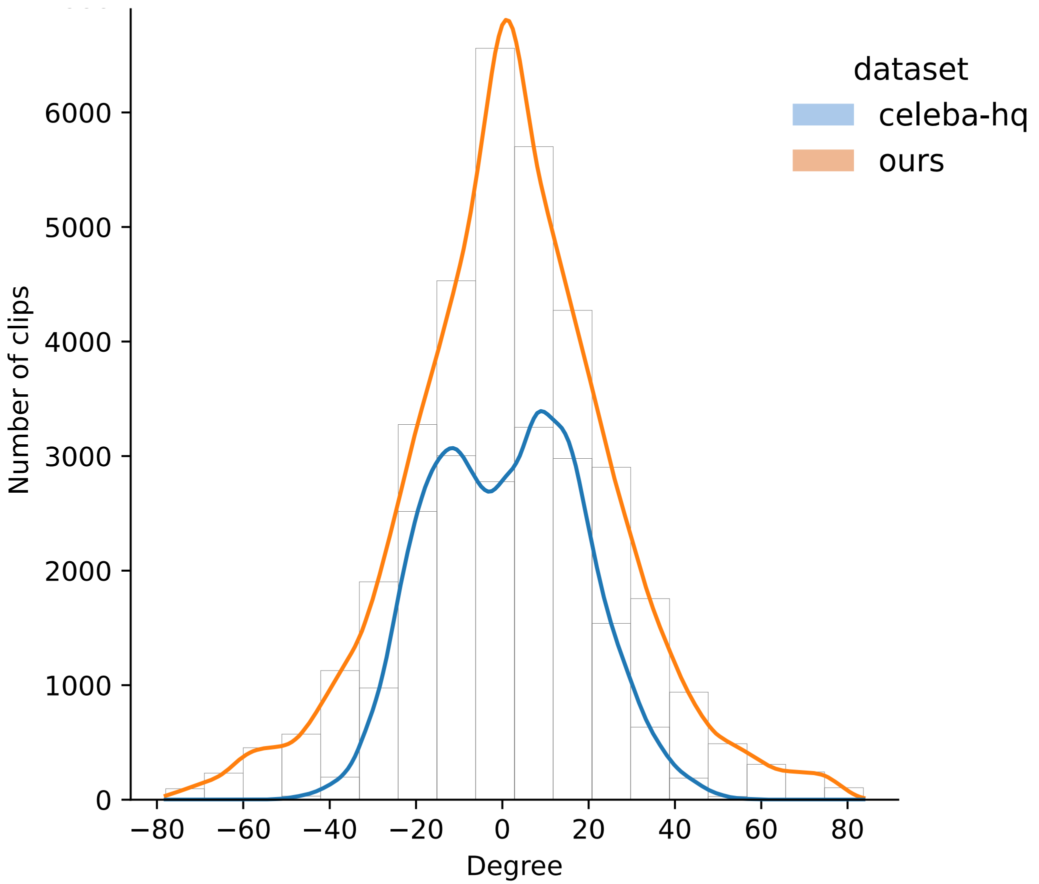
Video quality is measured by VSFA



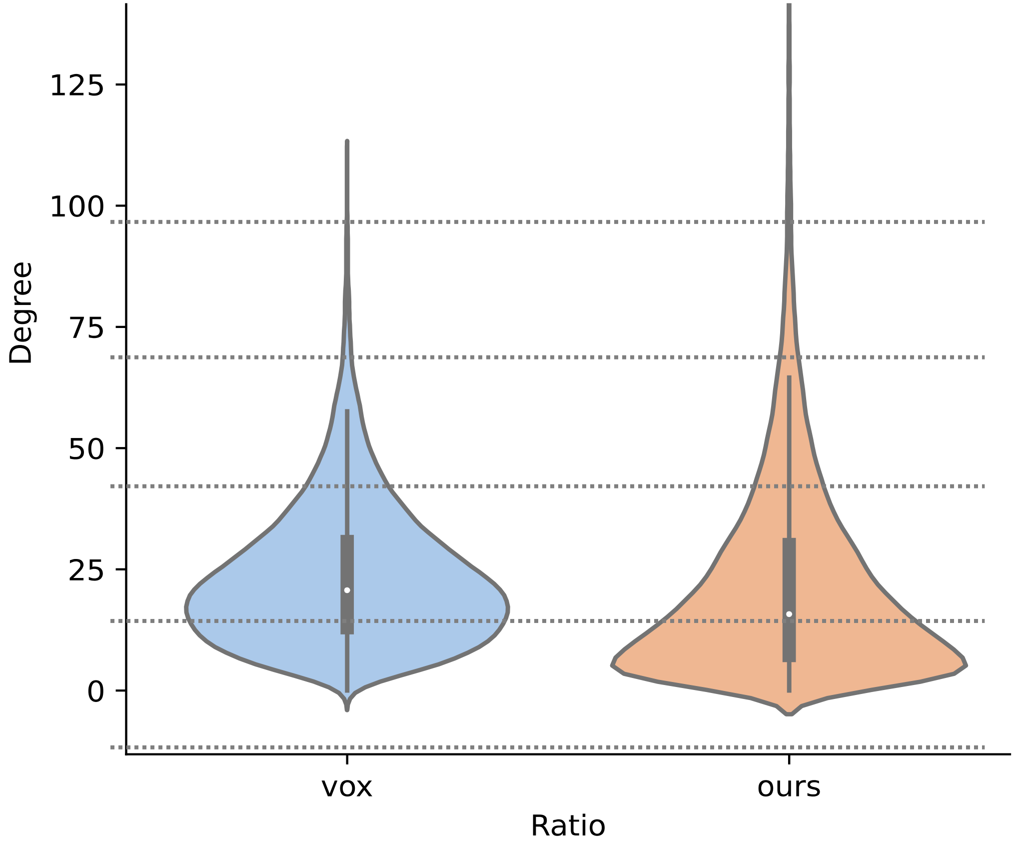
Distribution Comparison

VoxCeleb – Head Pose

CelebV-HQ is more diverse and smoother than VoxCeleb.



(a) Distribution of average pose

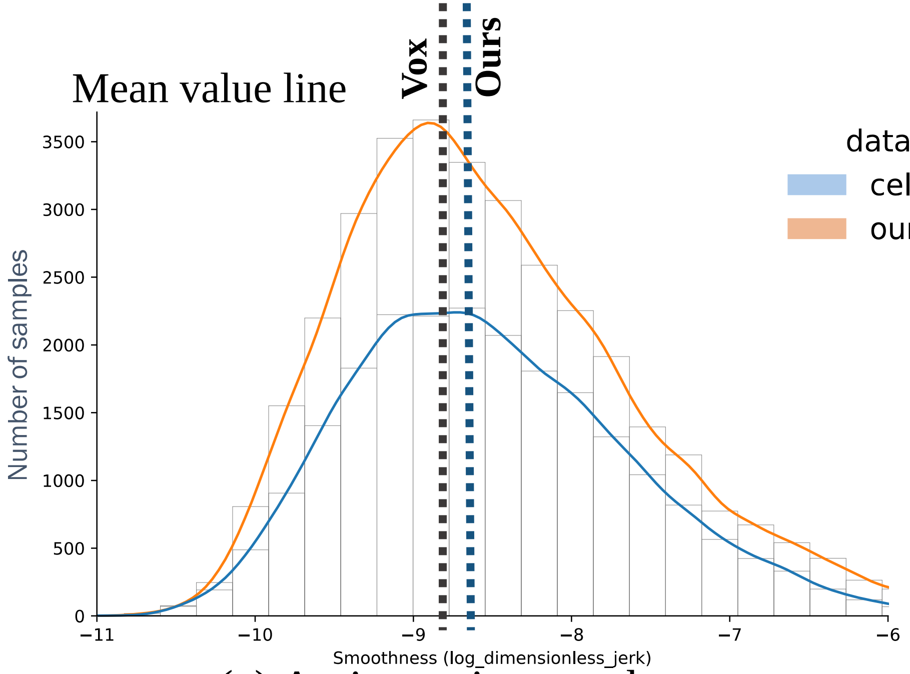


(b) Distribution of movement range

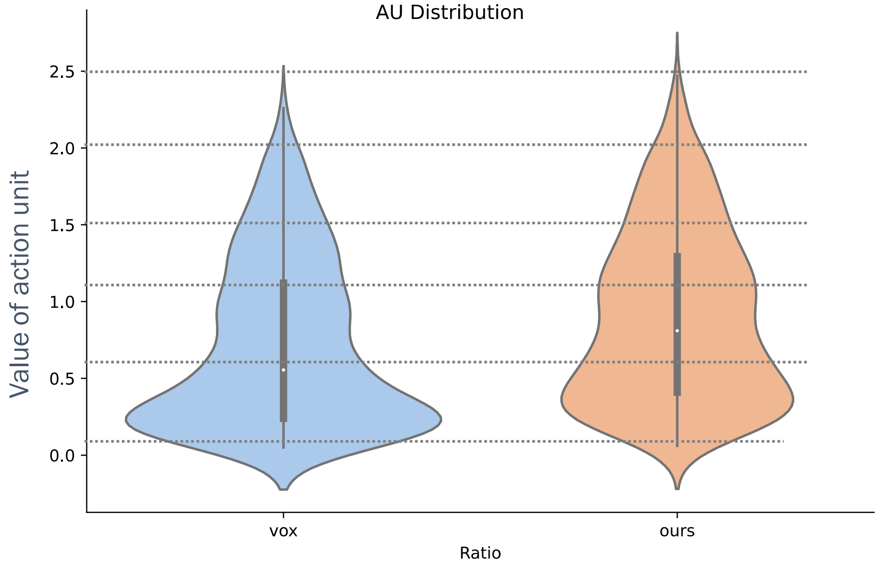
Distribution Comparison

VoxCeleb – Action Unit

CelebV-HQ is analyzed in both muscle movement naturalness and richness



(a) Action unit smoothness



(b) Action unit distribution

Benchmark

Unconditional Video Generation

VideoGPT

MoCoGAN-HD

DIGAN

StyleGAN-V

Table: FVD/FID Metrics Comparsion

	FaceForensics [65]		Vox [59]		MEAD [82]		CelebV-HQ	
	FVD (↓)	FID (↓)	FVD (↓)	FID (↓)	FVD (↓)	FID (↓)	FVD (↓)	FID (↓)
VideoGPT [90]	185.90	38.19	187.95	65.18	233.12	75.32	177.89	52.95
MoCoGAN-HD [75]	111.80	**7.12**	314.68	**55.98**	245.63	32.54	212.41	21.55
DIGAN [94]	62.50	19.10	201.21	72.21	165.90	43.31	72.98	19.39
StyleGAN-V [73]	**47.41**	9.45	**112.46**	60.44	**93.89**	**31.15**	**69.17**	**17.95**


Code and Models
CelebV-HQ:
A Large-Scale Video Facial Attributes Dataset
ECCV 2022

StyleGAN-Human:

A Data-Centric Odyssey of Human Generation

Jianglin Fu[1*], Shikai Li[1*], Yuming Jiang[2], Kwan-Yee Lin[1], Chen Qian[1], Chen Change Loy[2], Wayne Wu[1,3†], Ziwei Liu[2]

[1]SenseTime Research, [2]S-Lab, Nanyang Technological University, [3]Shanghai AI Laboratory

ECCV 2022 * Equal Contributions

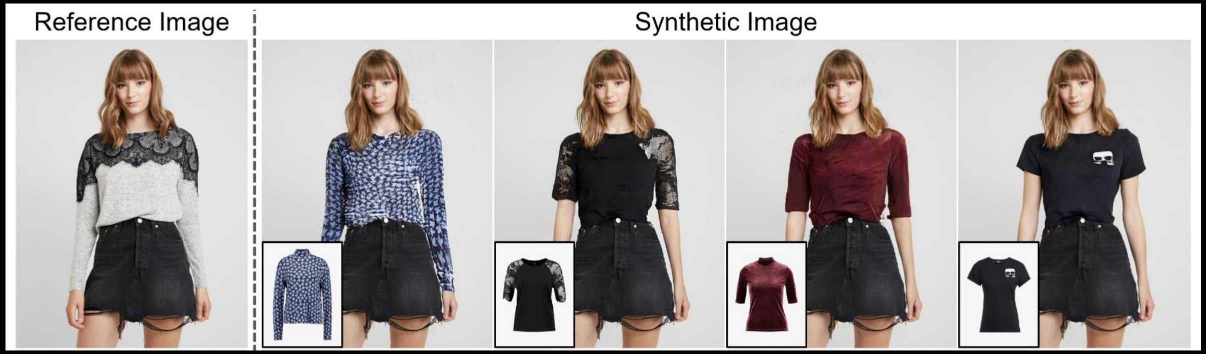
StyleGAN-Human:

A Data-Centric Odyssey of Human Generation

Introduction

Generating clothed humans

Virtual Try-on



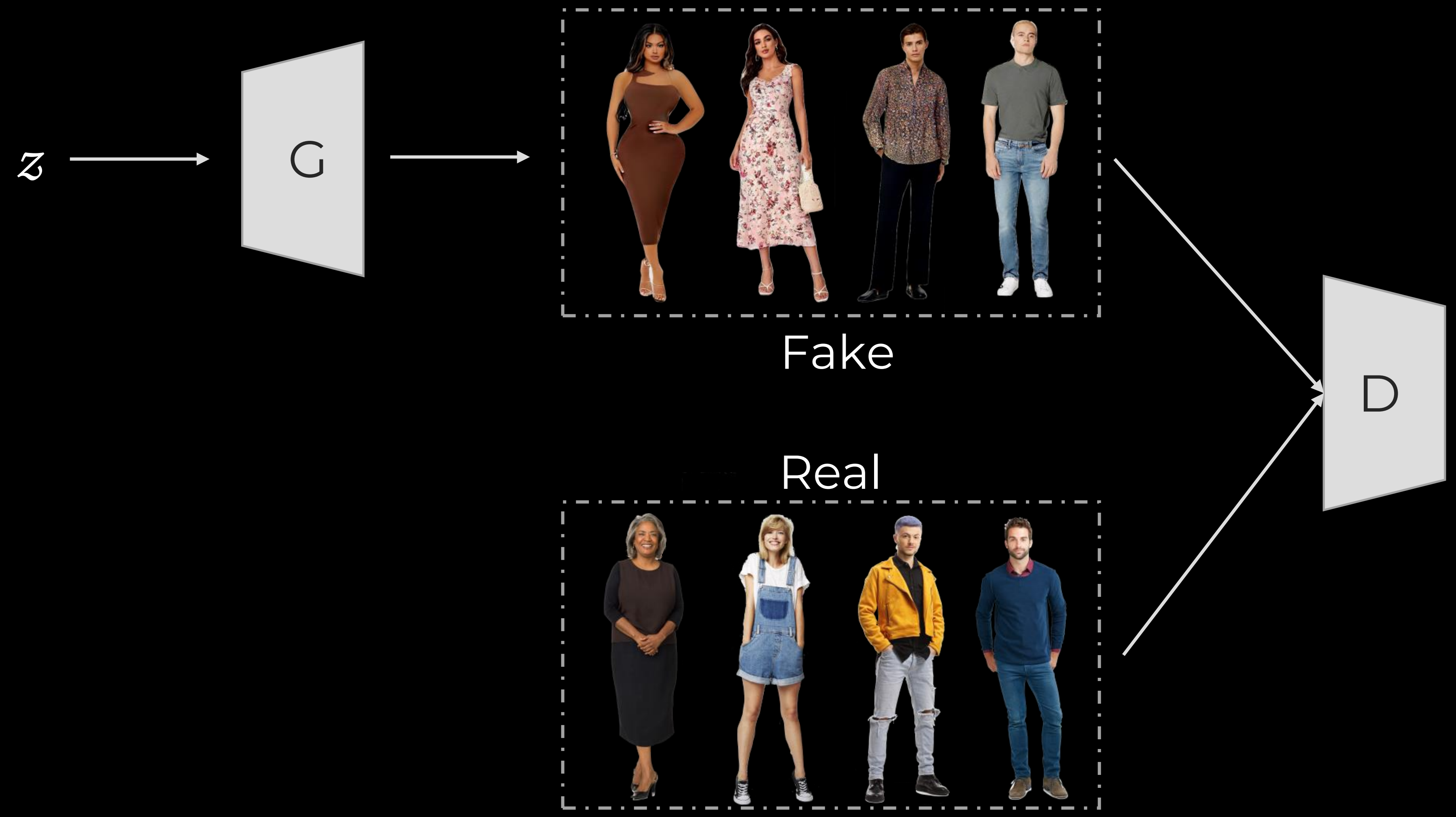
Viton-HD [Choi et al. 2021]

Human Motion Transfer

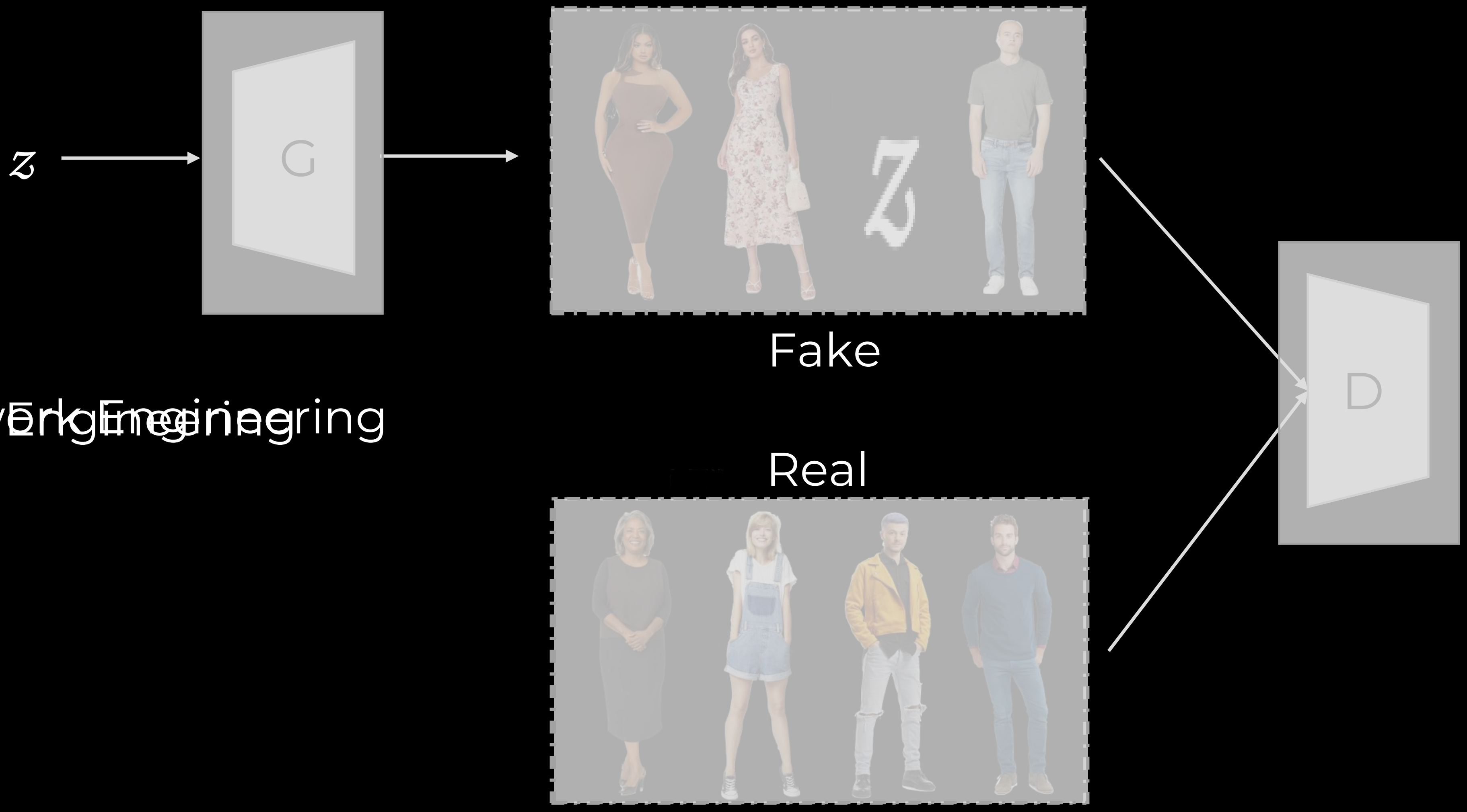
Liquid Warping GAN [Liu et al. 2019]

Generative Adversarial Networks

Unconditional Human Generation



Unconditional Human Generation



Network Engineering

Data Engineering

Compare with Public Dataset

Dataset	Image Number	Mean Resolution	**Labeled Attributes**	Full-Body Ratio
DeepFashion	146,680	1101x750	✓	6.8%
Market1501	32,668	128x64	✓	100%
ATR	7,700	400x600	✓	76%
LIP	50,462	197x345	✓	37%
VITON	16,253	256x192	✗	0%
Ours	**?**	**?**	**?**	**?**

Data Collection

From the Internet:

Images from Flickr with CC0 License

Images from Pixabay with Pixabay License

Images from Pexels with Pexels License

From the data providers:

Images from databases of individual photographers, modeling agencies and other suppliers .
(These images are internal used only and non-transferable)

Data Processing

Background

Resolution

Body Positon

Missing
Body-Part

Extreme Posture

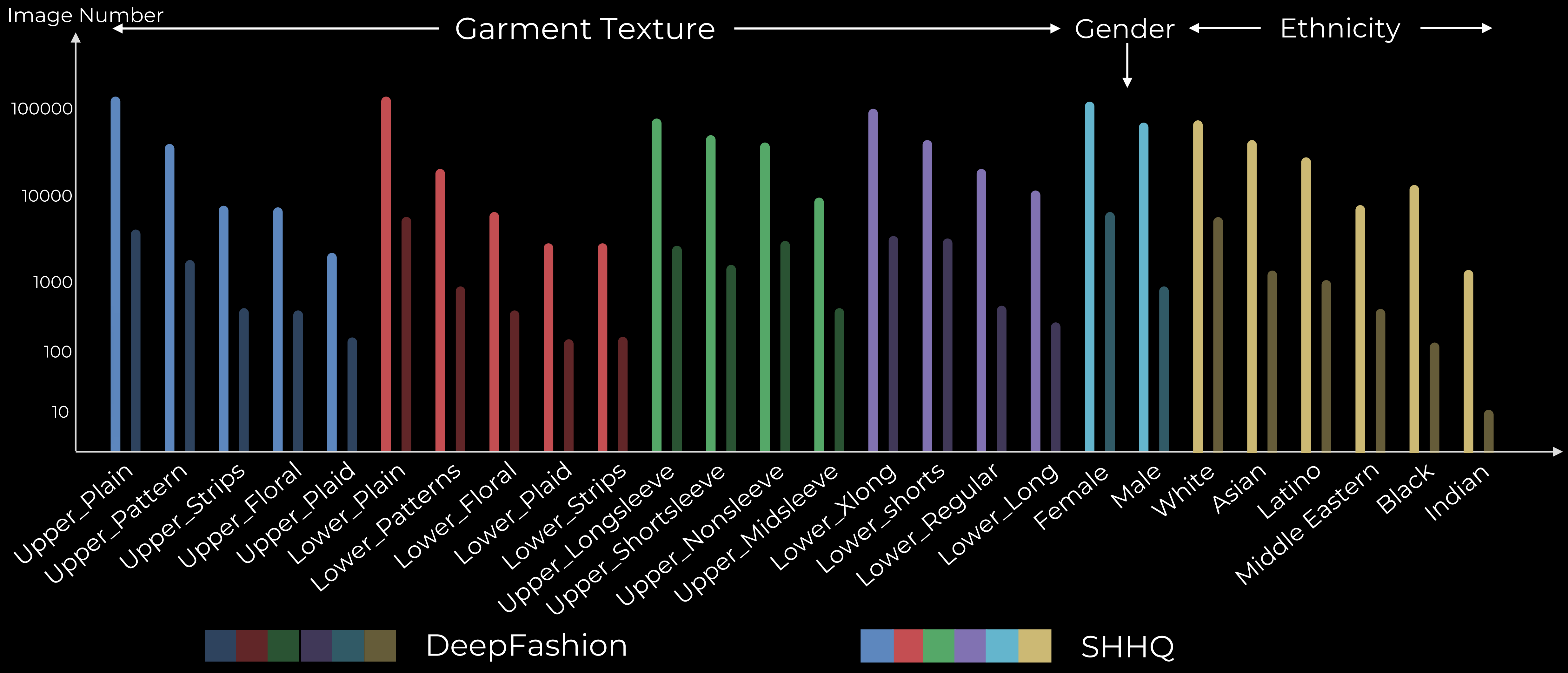
Multi-Person

Compare with Public Dataset

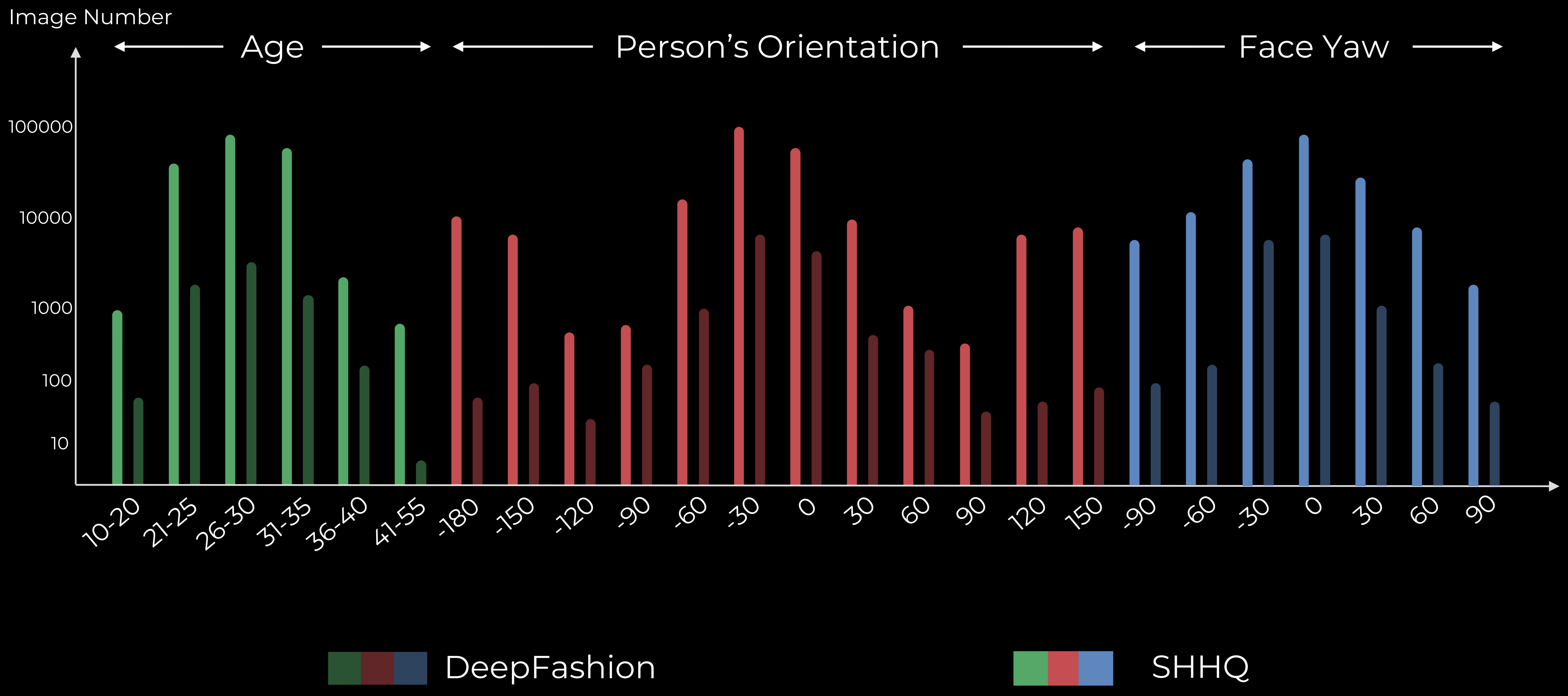
Dataset	Image Number	Mean Resolution	Labeled Attributes	Full-Body Ratio
DeepFashion	146,680	1101x750	✓	6.8%
Market1501	32,668	128x64	✓	100%
ATR	7,700	400x600	✓	76%
LIP	50,462	197x345	✓	37%
VITON	16,253	256x192	✗	0%
Ours	**231,176**	**1024x512**	✓	**100%**

Stylish-Humans-HQ (SHHQ)

Statistics of collected dataset



Statistics of collected dataset

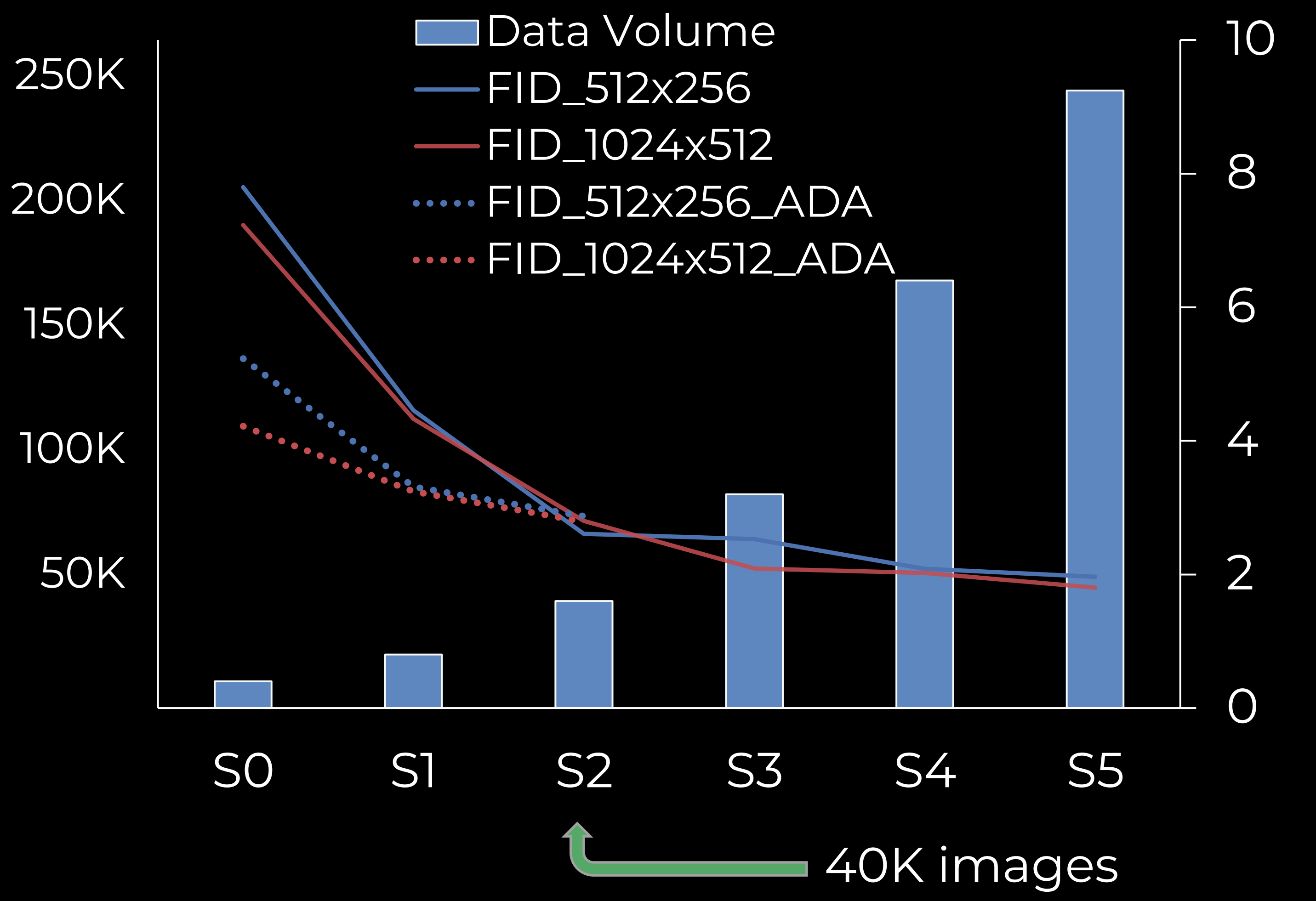


Question-1:
What is the relationship between the data size and the generation quality?

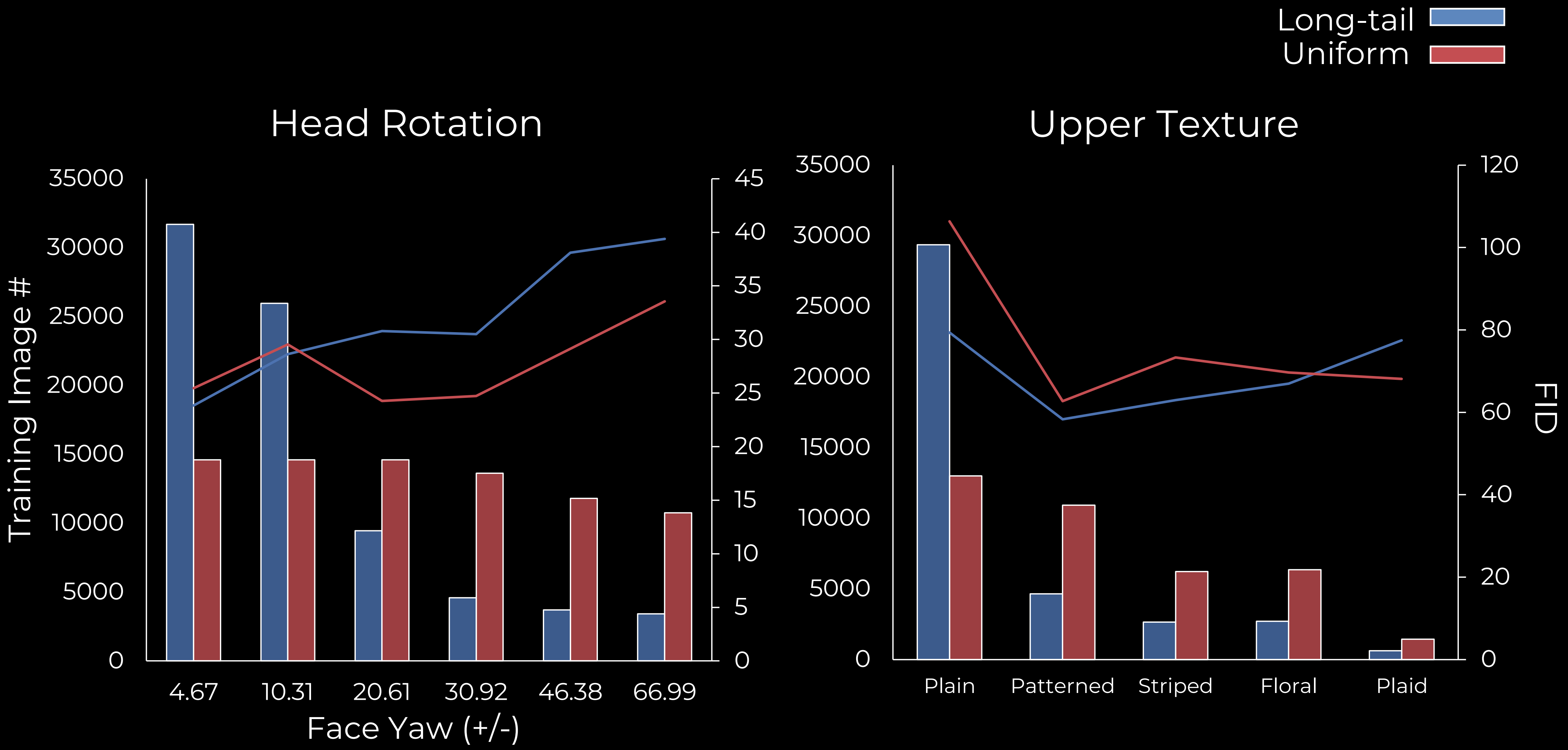
Question-2:
What is the relationship between the data distribution and the generation quality?

Question-3:
What is the relationship between the scheme of data alignment and the generation quality

Experiments: Data Volume



Experiments: Data Distribution



Experiments: Data Pre-processing

~~Center 1: center of face bbox~~

~~Center 2: position of pelvis~~

Center 3: middle point of body

Model Zoo

Face

StyleFlow
Editing on pose

InterFaceGAN
Editing on gender

StyleNerf
Preserve 3D consistency

StyleGAN | StyleGAN2 | StyleGAN3

Human

?

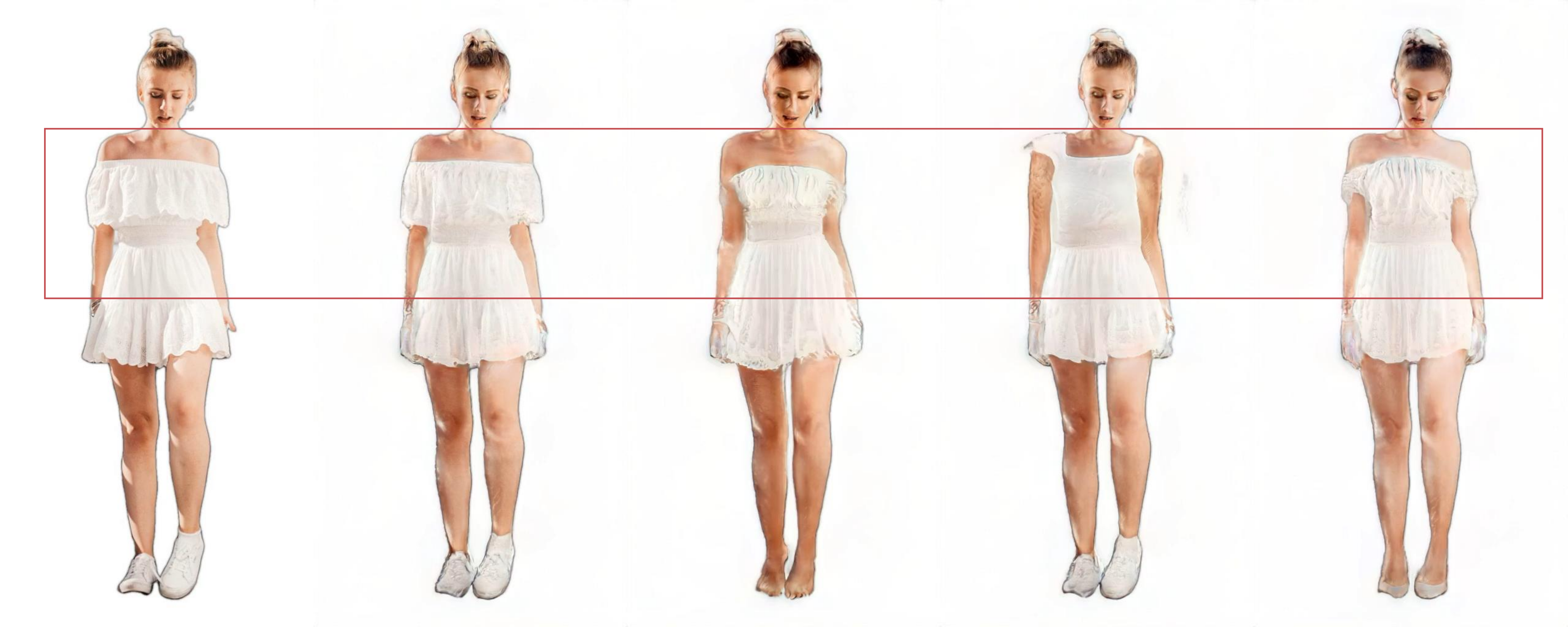
Baseline Results



Editing Benchmark

Source InterfaceGAN StyleSpace SeFa

Source InterfaceGAN StyleSpace SeFa



Real Image PTI Inversion InterfaceGAN StyleSpace SeFa

Real Image PTI Inversion InterfaceGAN StyleSpace SeFa

Style
Mixing


Source
Refenrece

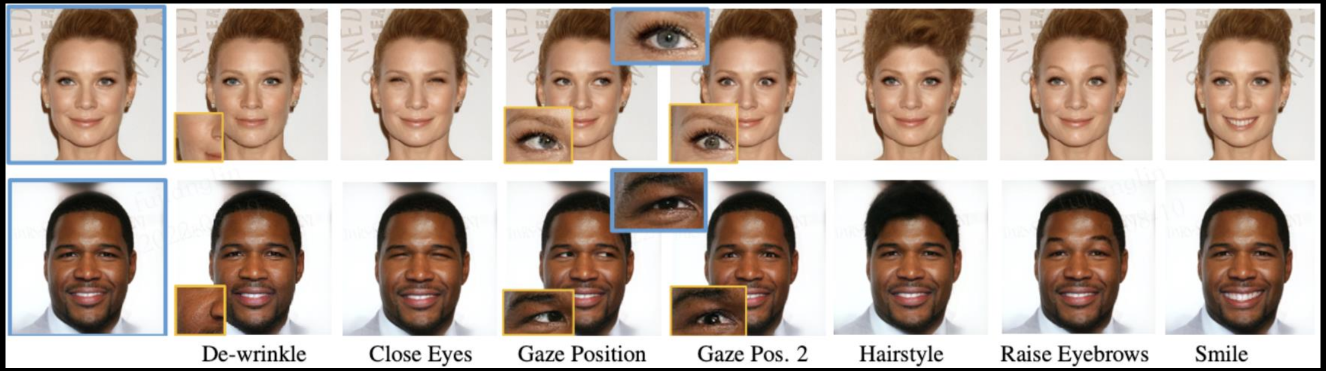

Iter:1
Optimize face
Iter:1
Optimize face
Iter:1
Optimize face
InsetGAN Results

SHHQ-1.0

1. Images obtained from the Internet (Flickr, Pixabay, Pexels).
2. Processed 9991 DeepFashion images (retain only full body images).
3. 1940 African images from the InFashAI dataset to increase data diversity.

Future Work

1. Human Generation / Editing

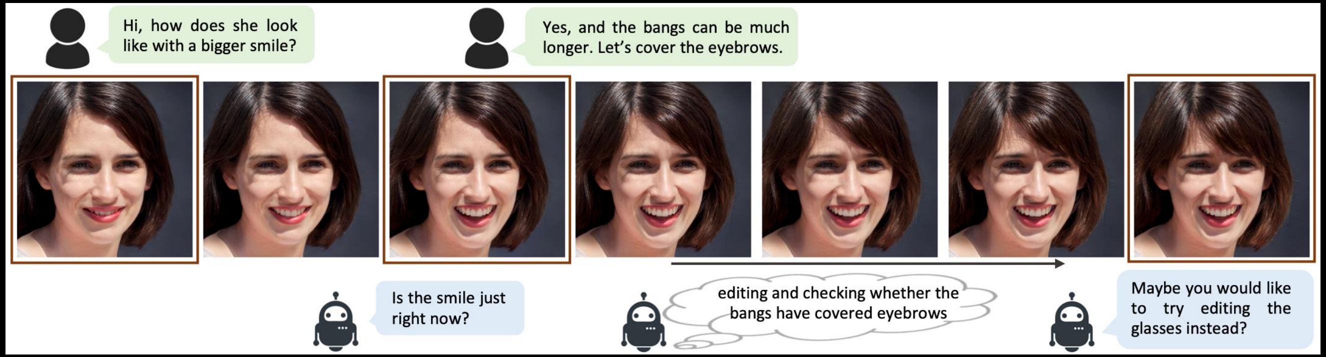


EditGAN [Ling et al. 2021]

2. Neural Rendering

StyleNerf [Gu et al. 2022]

3. Multi-modal Generation



Talk-to-Edit [Jiang et al. 2021]

Code and Models

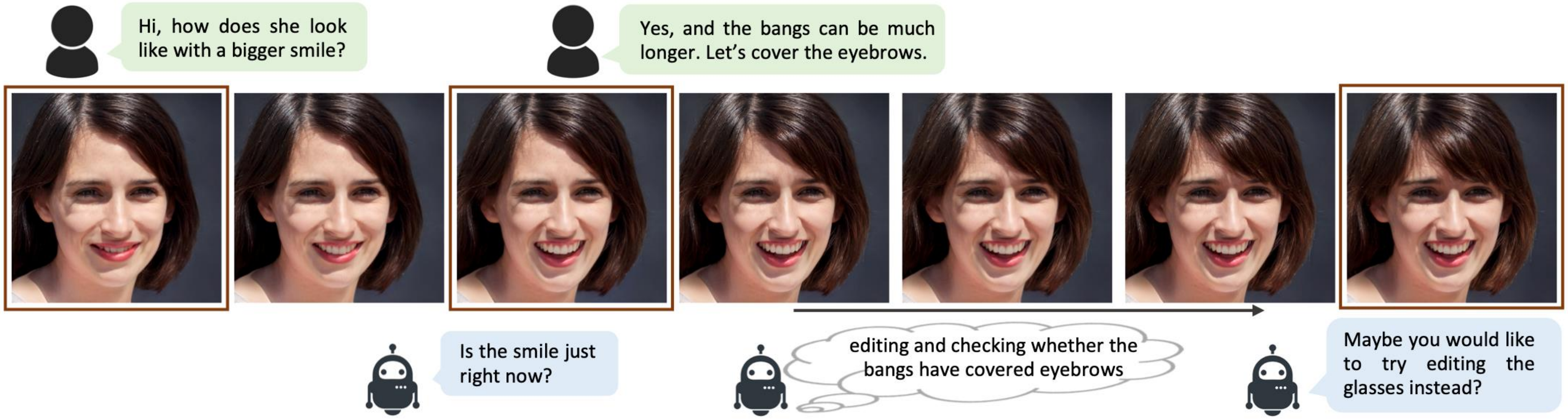
Interactive Generative Models

Talk-to-Edit: Fine-Grained Facial Editing via Dialog

Yuming Jiang[1*] Ziqi Huang[1*] Xingang Pan[2] Chen Change Loy[1] Ziwei Liu[1✉]

[1] S-Lab, Nanyang Technological University [2] The Chinese University of Hong Kong

Talk-to-Edit



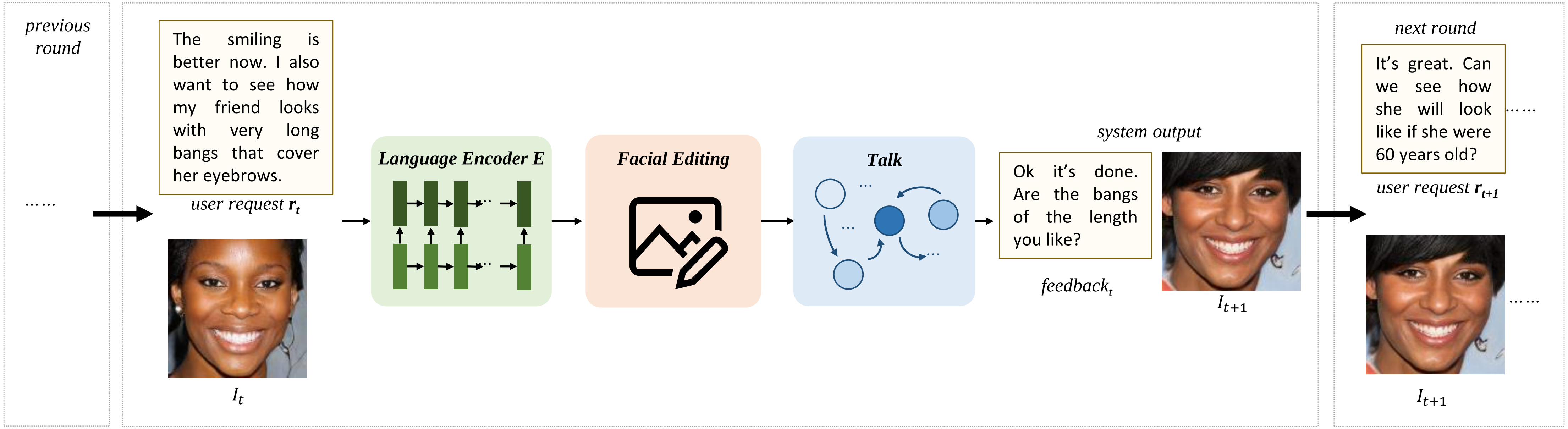
- Propose to perform fine-grained facial editing via dialog
- Propose to model a location-specific semantic field in GAN latent space
- Achieve superior results with better identity preservation and smoother change
- Contribute a large-scale visual-language dataset CelebA-Dialog

Motivation

- Facial Editing:
  - enable users to manipulate facial images in their desired way
- Current Facial Editing Systems
  - image-to-image translation models: do not allow controls
  - fixed interaction ways:
    - semantic segmentation map, a reference image, a sentence describing a desired effect
- Dialog-based Facial Editing
  - natural language is a flexible interaction way for users
  - system can provide feedback
  - editing is performed round by round via dialog

Talk-to-Edit Pipeline

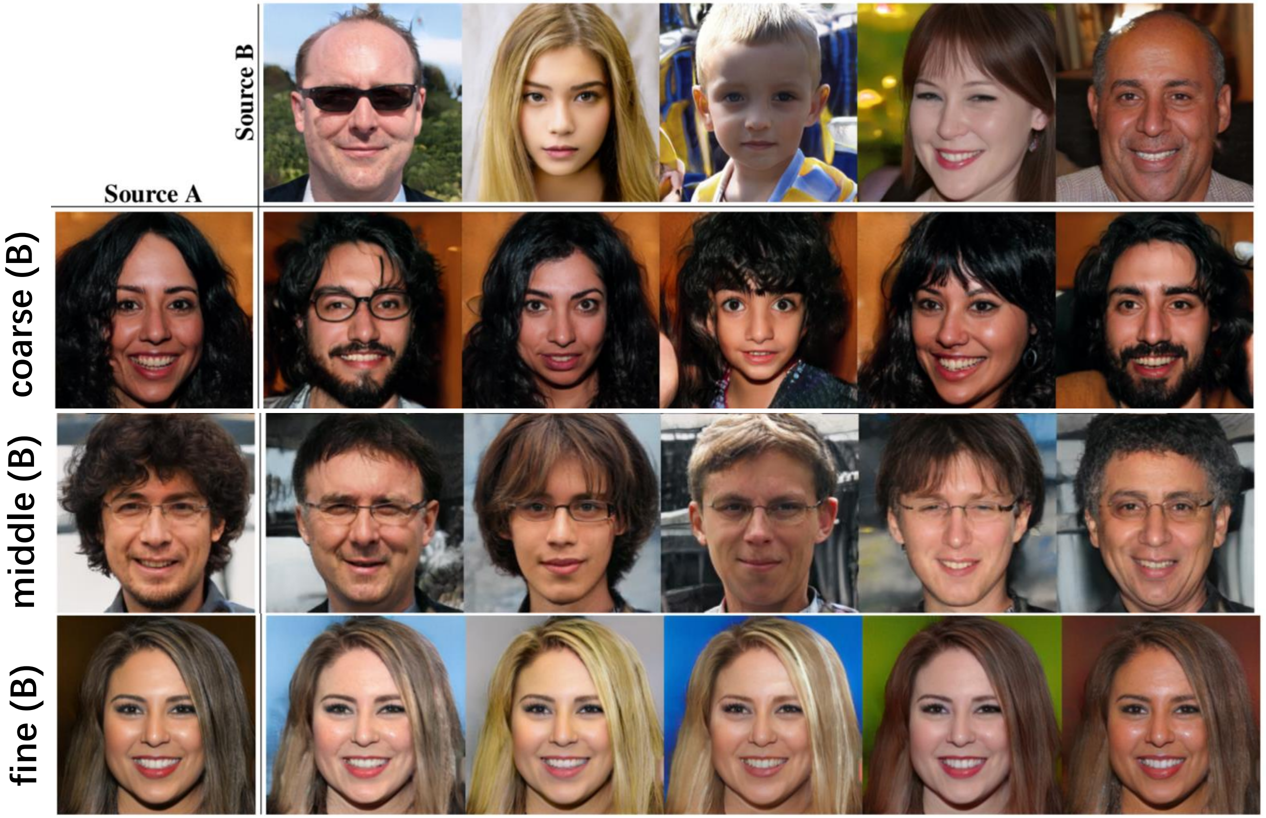
- Language Encoder: understands user request
- Facial Editing: performs facial editing according to the language request
- Talk Module: provides meaningful natural language feedback



Facial Editing Module

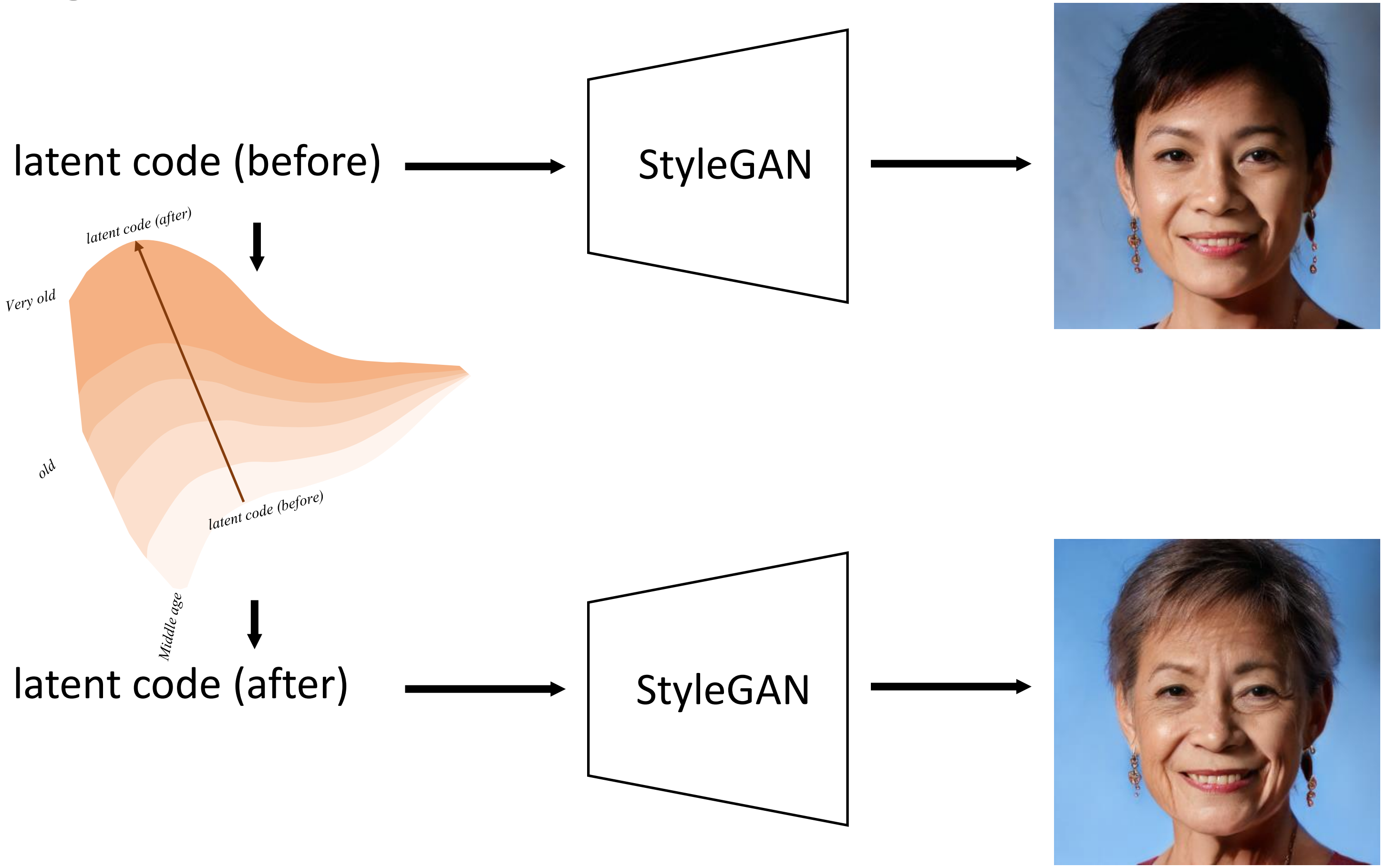
- Interactions by dialog
  - users may change their thoughts during editing
  - tuning an overly laughing face back to a moderate smile
- Continuous and fine-grained facial editing
- Using Pretrained StyleGAN as the face generator

StyleGAN



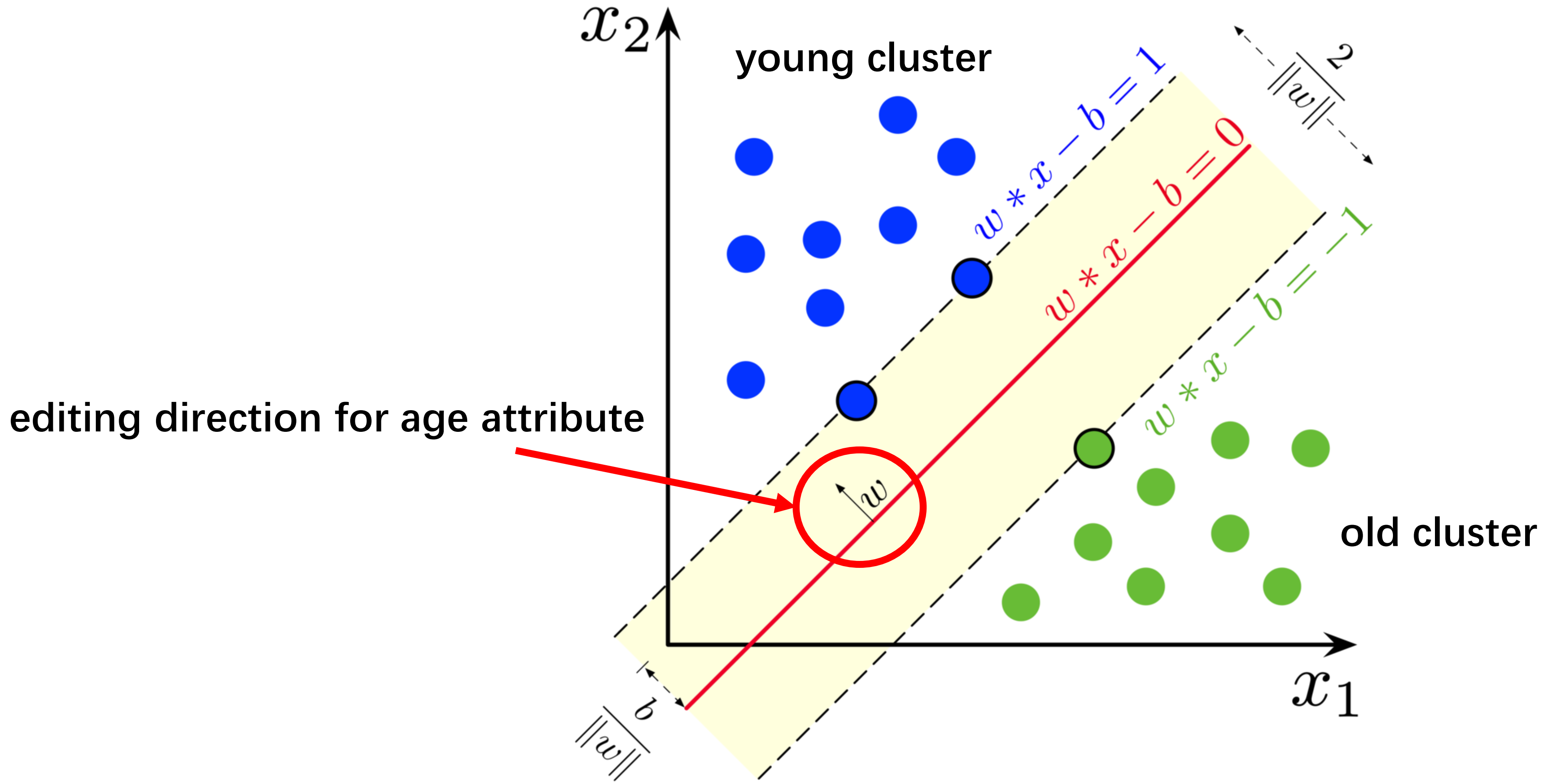
Editing in GAN Latent Space

- Existing latent based methods

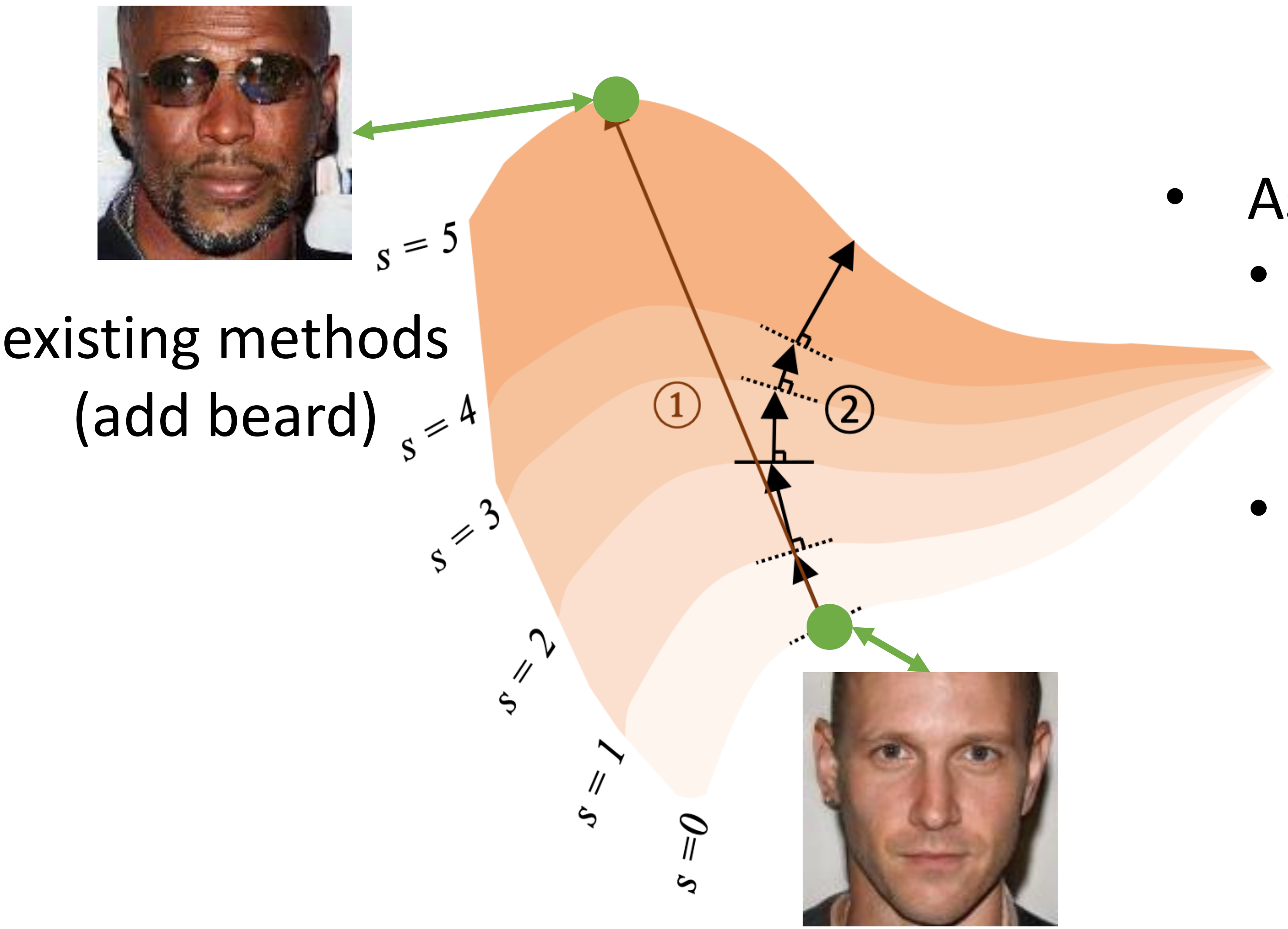


InterFaceGAN

- Train an SVM to find the editing direction for the target attribute

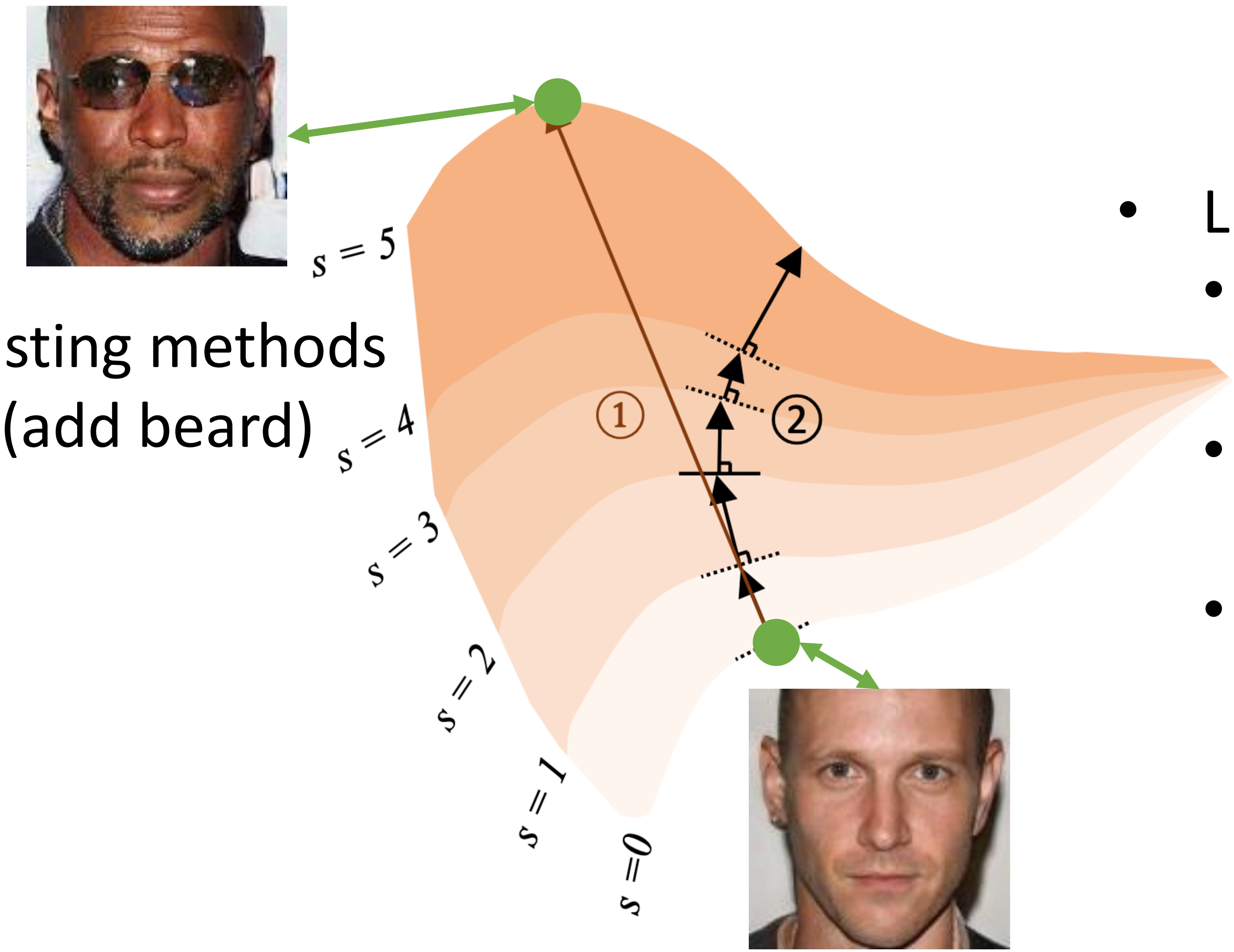


Editing in GAN Latent Space



- Assumptions of existing methods
  - The attribute change is achieved by traversing along a straight line
  - Different identities share the same latent directions

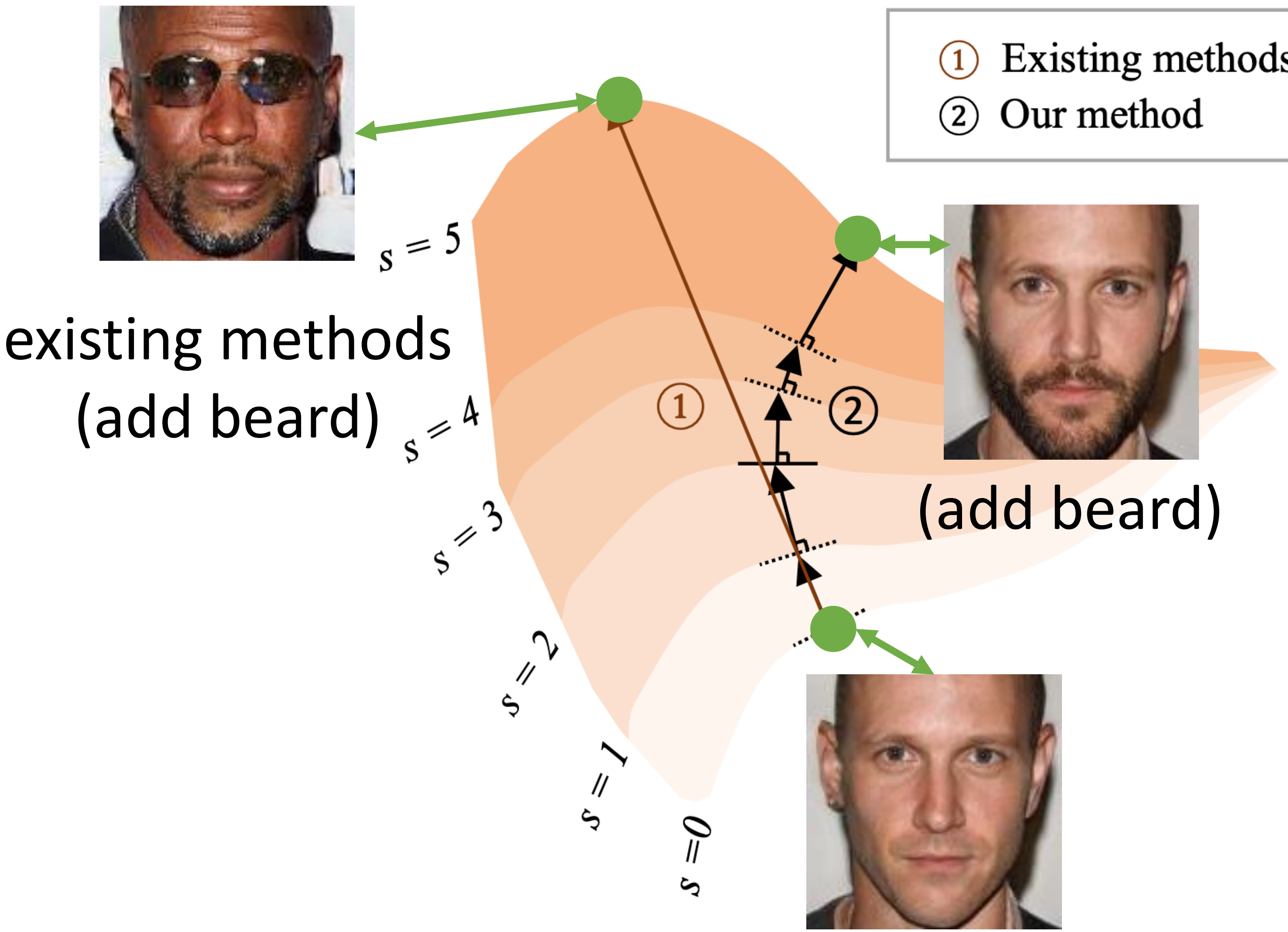
Editing in GAN Latent Space



existing methods
(add beard)

- Limitations of existing methods
  - The identity would drift during editing
  - Other irrelevant attributes would be changed as well
  - Artifacts would appear

Semantic Field in GAN Latent Space



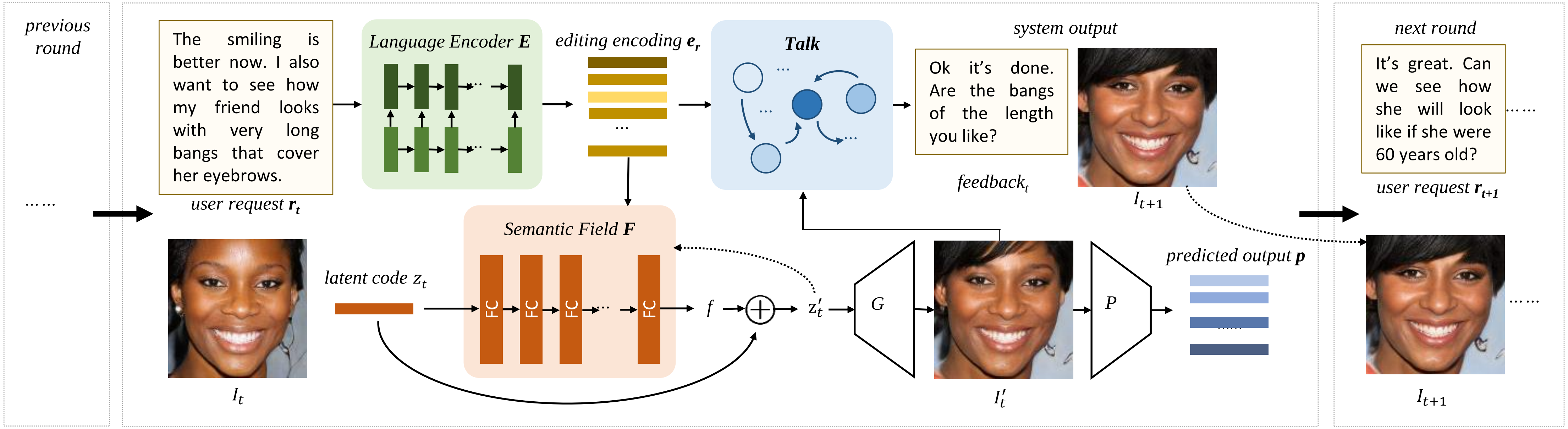
- Semantic field:
  Consider the non-linearity of the attribute transition
- Ours: move the latent code along the curved field line

$$s_a + \int_{\boldsymbol{z}_a}^{\boldsymbol{z}_b} \boldsymbol{f}_z \cdot d\boldsymbol{z} = s_b$$

- Ours: smoother change and better identity preservation

Talk-to-Edit Pipeline

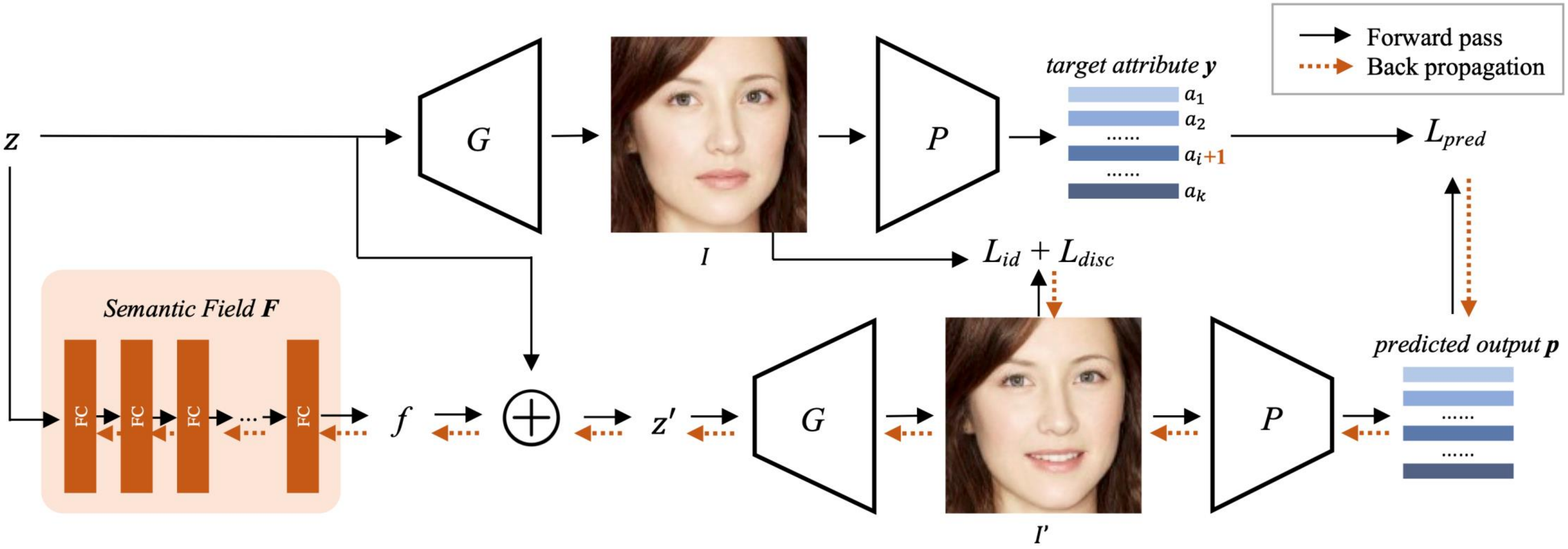
- Language Encoder: understands user request
- Semantic Field: performs fine-grained editing
- Talk Module: provides meaningful natural language feedback



Semantic Field Training

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
S-LAB FOR ADVANCED INTELLIGENCE

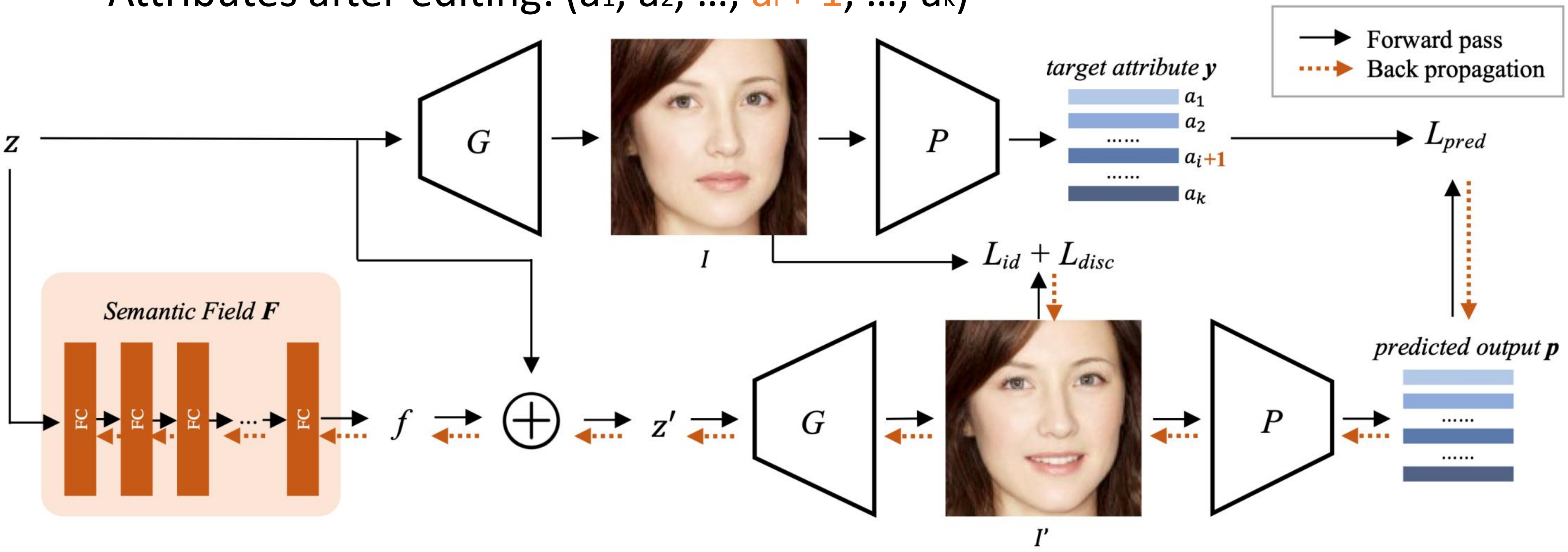
- Predictor Loss: change desired attribute, keep irrelevant attributes
- Identity keeping loss: preserve identity
- Discriminator loss: ensure photo-realism



Semantic Field Training

- **Predictor Loss**: change desired attribute, keep irrelevant attributes
  - For one attribute, degrees are classified into 6 fine-grained levels.
  - Original attributes: (a_1, a_2, …, a_i, …, a_k)
  - Attributes after editing: (a_1, a_2, …, $a_i + 1$, …, a_k)

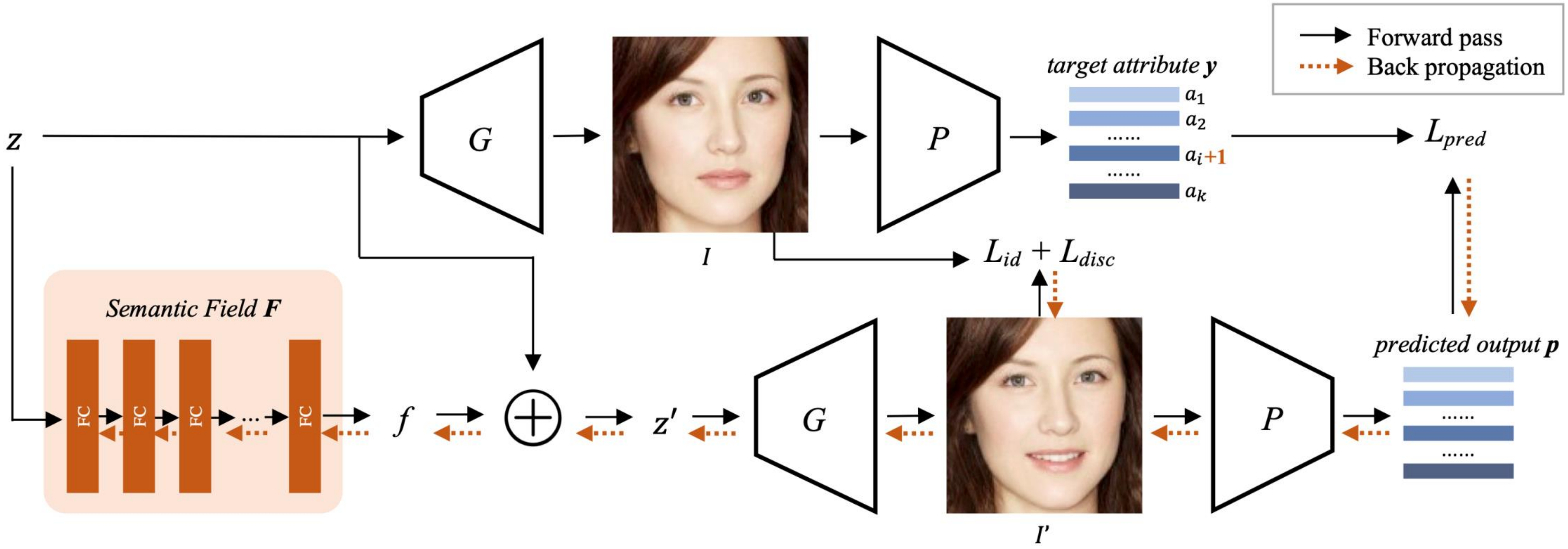
$$L_{pred} = -\sum_{i=1}^{k}\sum_{c=0}^{C} y_{i,c} log(p_{i,c}),$$



Semantic Field Training

- Identity keeping loss: preserve identity
  - Employ an off-the-shelf pretrained face recognition model to extract discriminative features

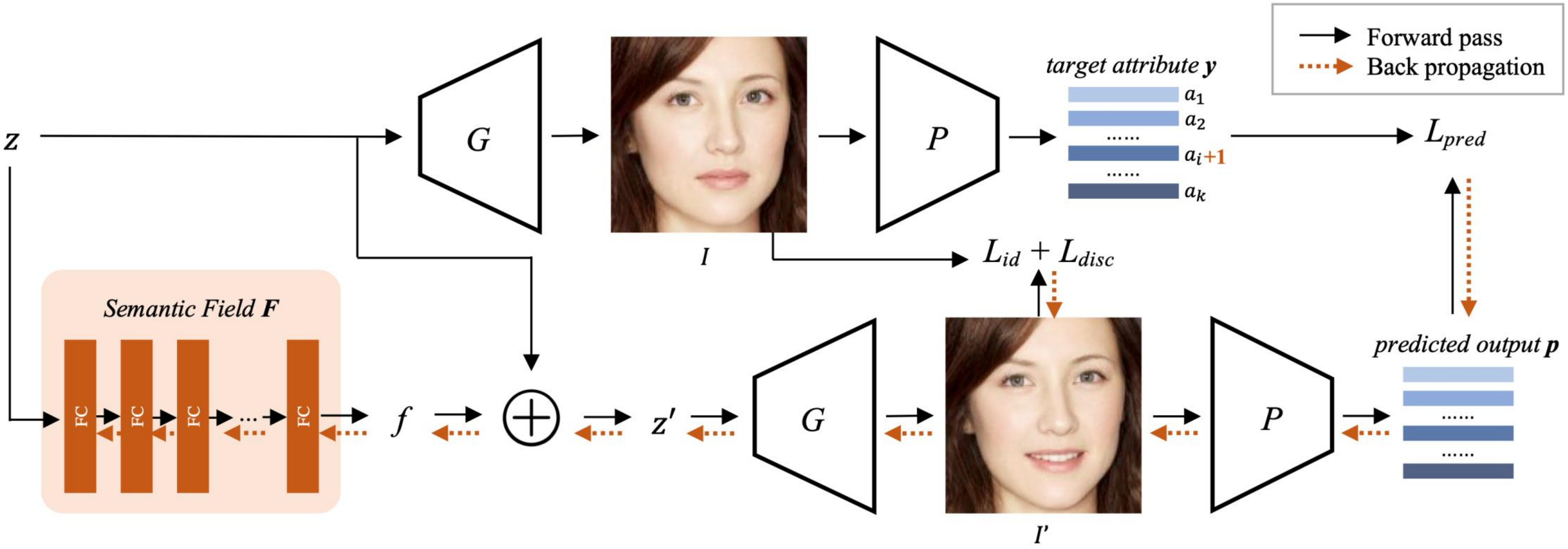
$$L_{id} = \|Face(\mathbf{I}') - Face(\mathbf{I})\|_1 \,,$$



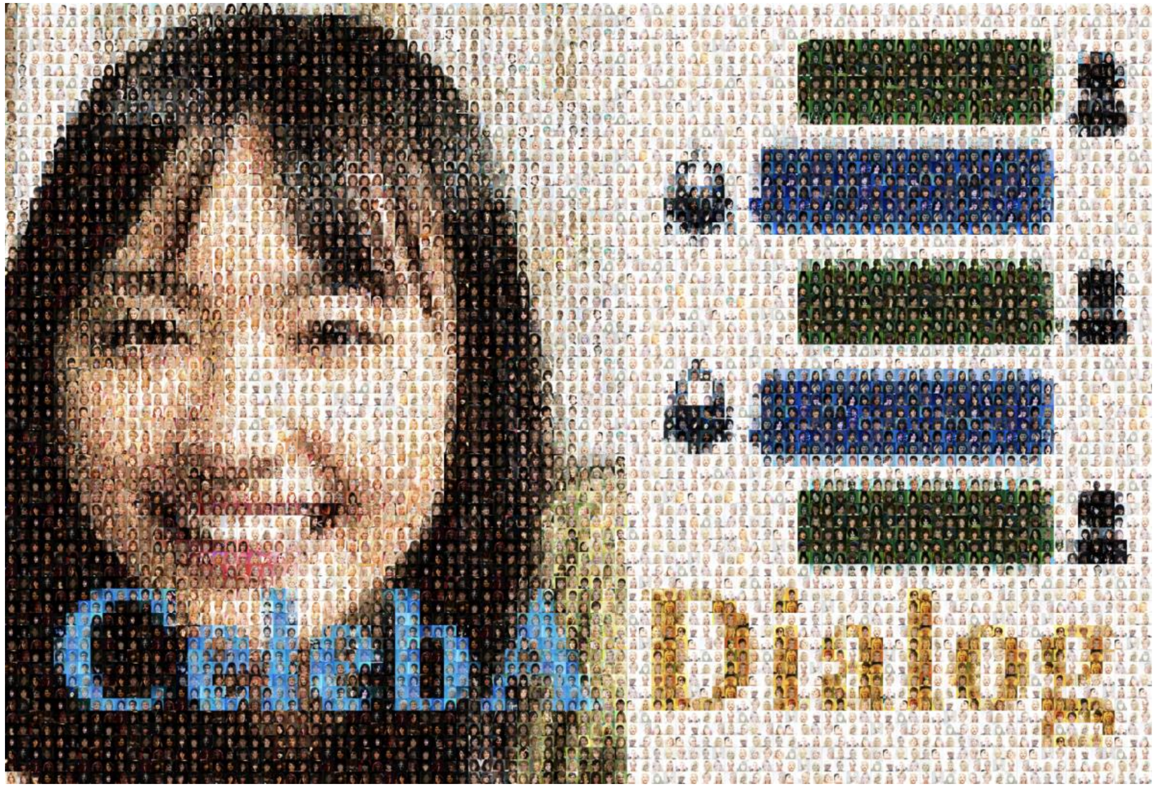
Semantic Field Training

- Discriminator loss: ensure photo-realism
  - Use the pretrained discriminator D coupled with the face generator

$$L_{disc} = -D(\mathbf{I}').$$



CelebA-Dialog Dataset

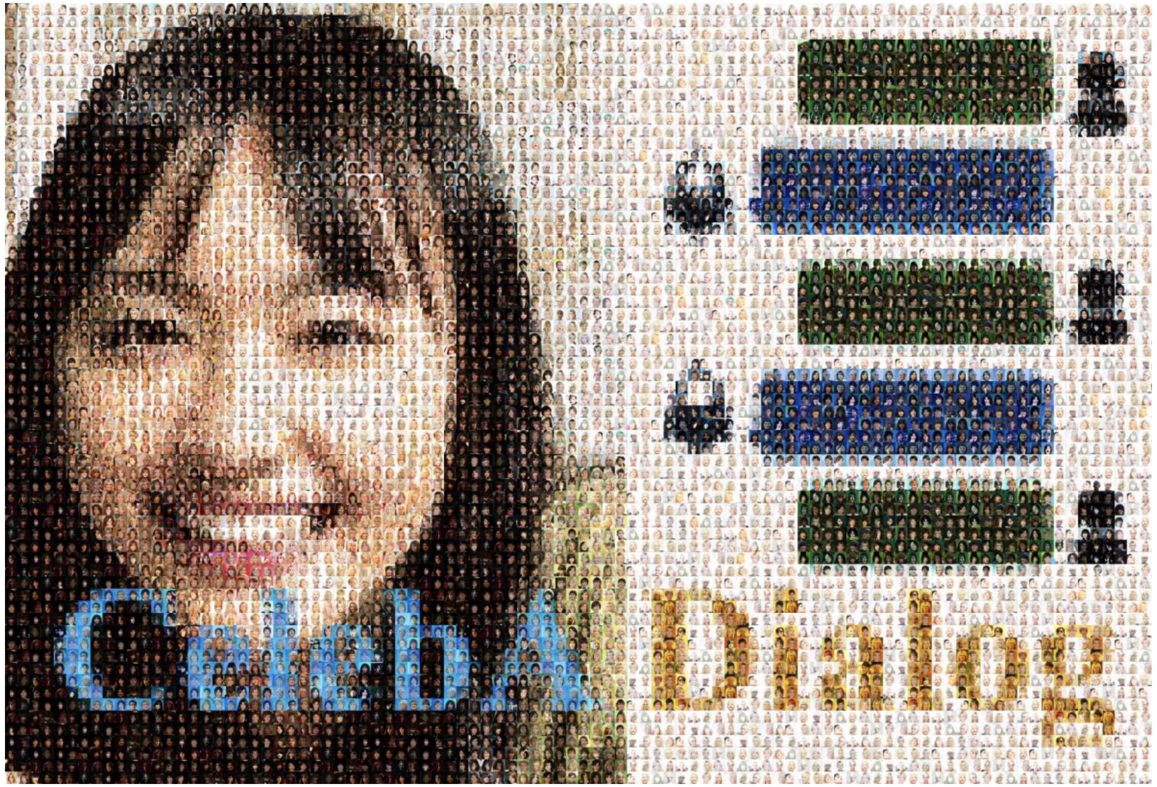


- Provide fine-grained attribute labels for attribute classifier training
- Languages for the training of language encoder and decoder

CelebA-Dialog Dataset

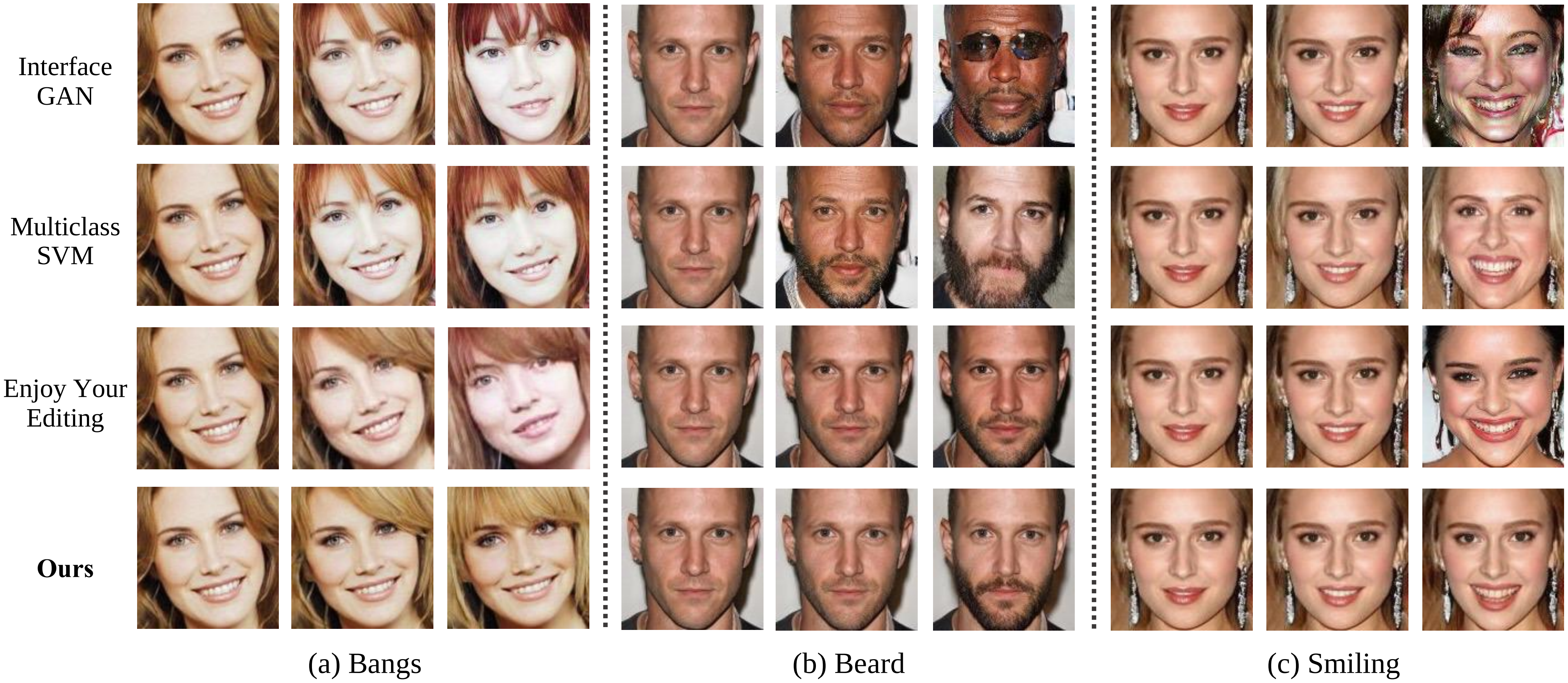
Attribute Degree	Fine-Grained Definition	Examples	
0	without bangs, full forehead exposed		*The lady has no bangs.*
1	very short bangs, 80% forehead exposed		*She has very short bangs covering her forehead.*
2	short bangs, 60% forehead exposed		*The man has short bangs that cover a small portion of the forehead.*
3	medium bangs, 40% forehead exposed		*The woman has bangs of medium length.*
4	long bangs, 20% forehead exposed		*The guy has long bangs.*
5	extremely long bangs, all forehead covered		*The woman has bangs that cover the eyebrows.*

CelebA-Dialog Dataset



- Large-scale visual-language dataset
- 202,599 face images
- Rich fine-grained labels (6 levels)
- Image captions describing attributes
- User editing requests
- Enable various tasks

Experimental Results



Experimental Results

- Talk-to-Edit preserves identity and irrelevant attributes better
- (Identity / Attribute) preservation score, both the lower the better

Methods	Bangs	Eyeglasses	Beard	Smiling	Young
InterfaceGAN	0.7621 / 0.7491	0.7831 / 1.1904	1.0213 / 1.6458	0.9158 / 0.9030	0.7850 / 1.4169
Multiclass SVM	0.7262 / 0.5387	0.6967 / 0.9046	1.1098 / 1.7361	0.7959 / 0.8676	0.7610 / 1.3866
Enjoy Your Editing	0.6693 / 0.4967	0.7341 / 0.9813	0.8696 / 0.7906	0.6639 / 0.5092	0.7089 / 0.5734
Talk-to-Edit (Ours)	0.6047 / 0.3660	**0.6229** / 0.7720	0.8324 / 0.6891	0.6434 / 0.5028	0.6309 / 0.4814
Talk-to-Edit (Ours) *	**0.5276 / 0.2902**	0.6670 / **0.6345**	**0.7634 / 0.5425**	**0.4580 / 0.3573**	**0.6234 / 0.2731**

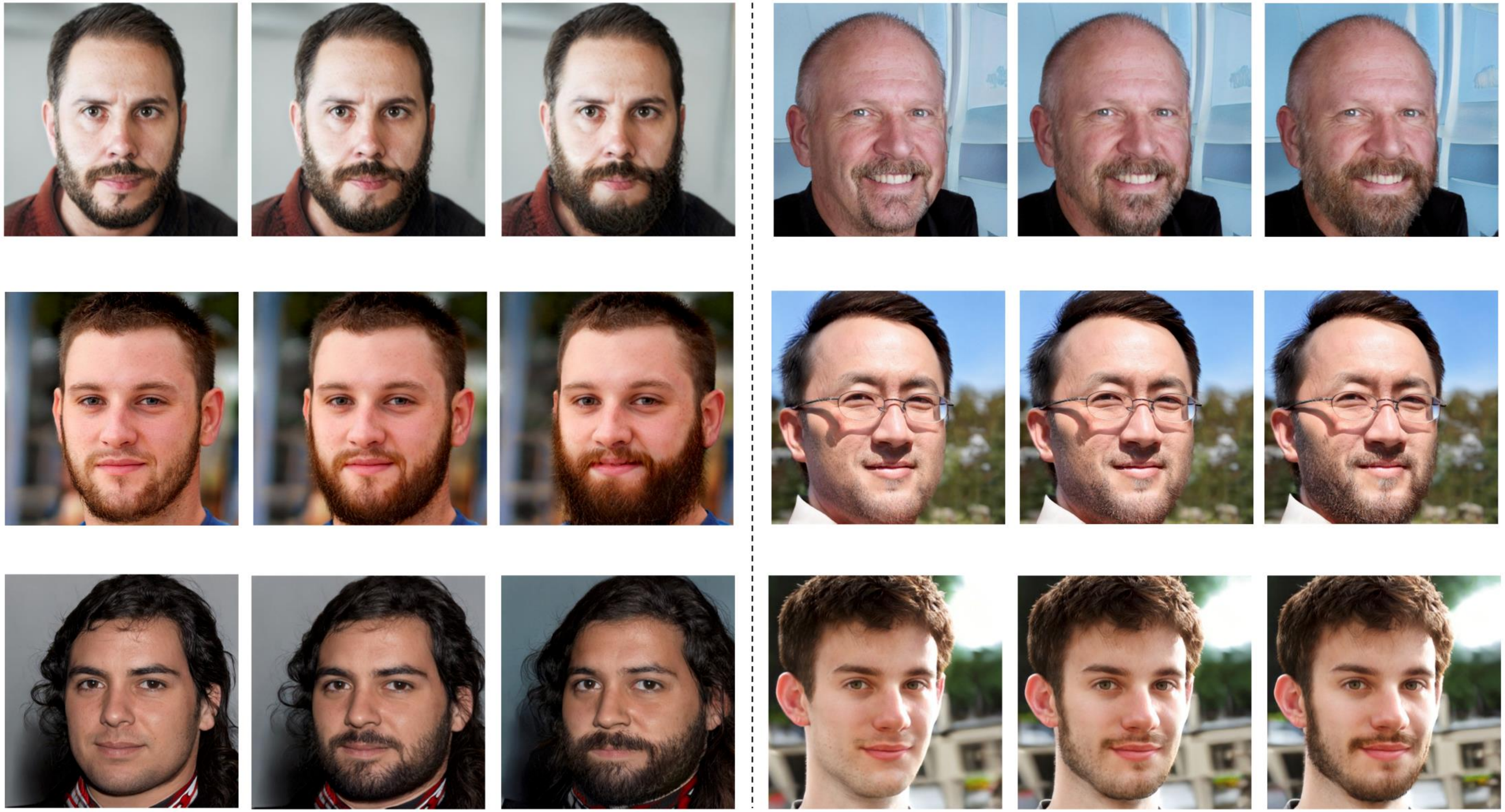
Experimental Results

(a) Bangs

Experimental Results

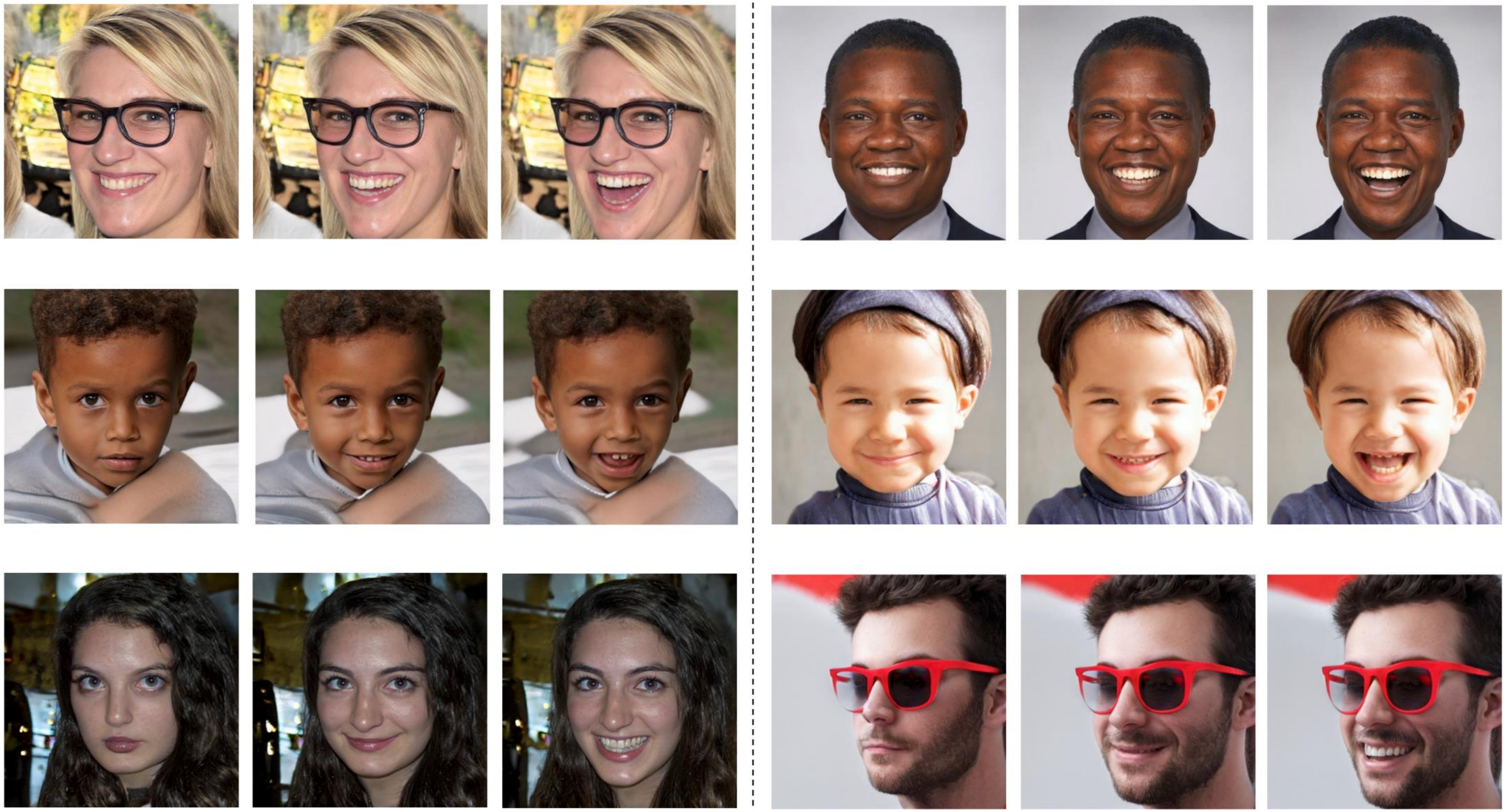
(b) Eyeglasses

Experimental Results



(c) Beard

Experimental Results

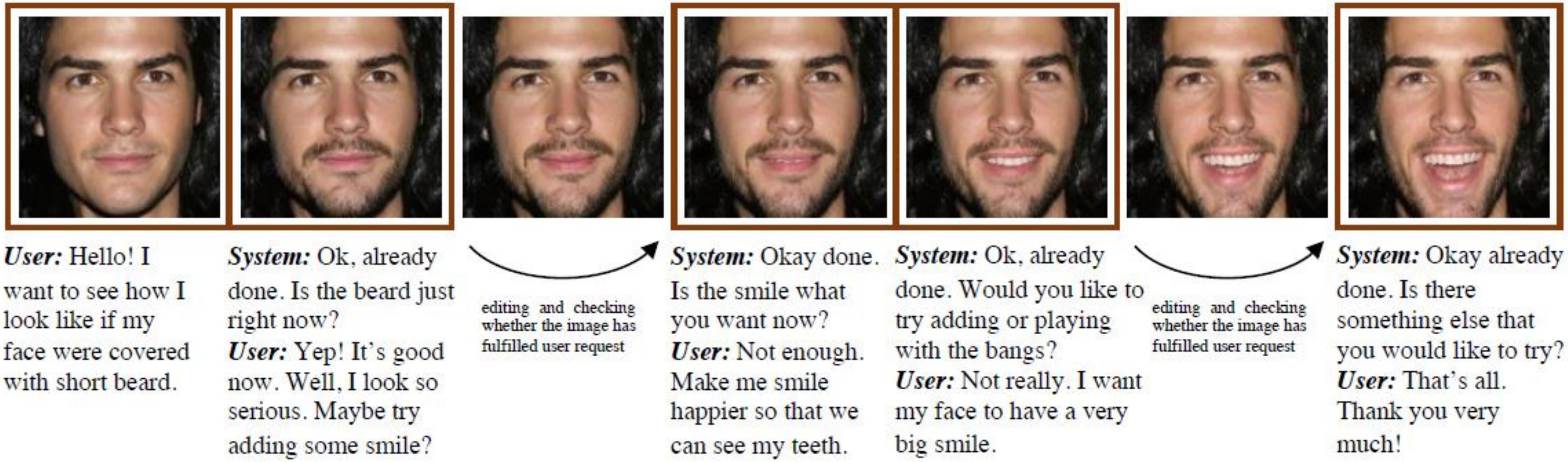


(d) Smiling

Experimental Results

(e) Young

Dialog based facial editing

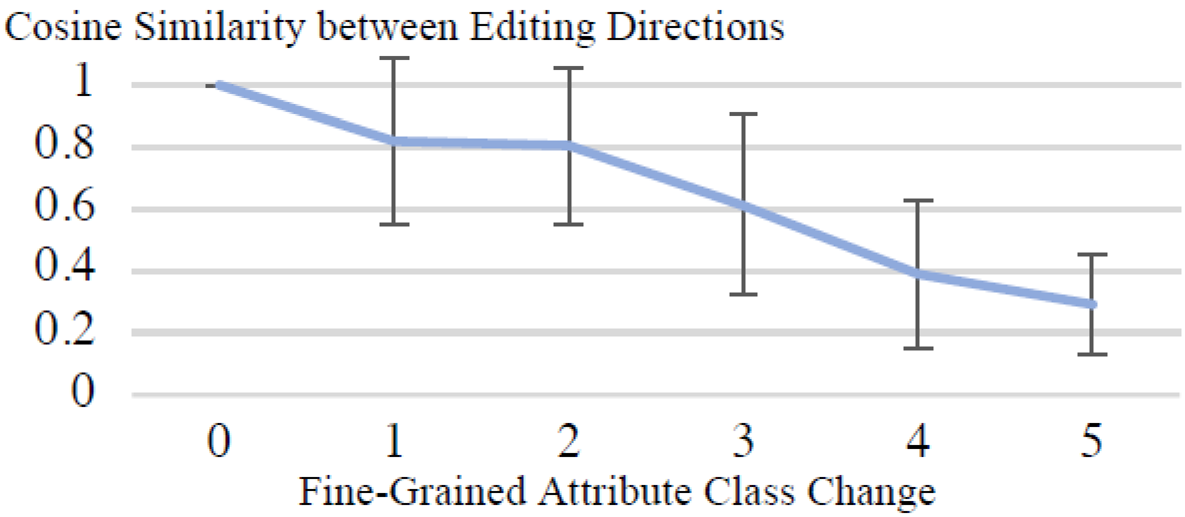


Editing in Real Images

- GAN inversion
  - Find the corresponding latent code z for real images in latent space
  - Finetune the latent code z as well as the weight of the StyleGAN

Further Analysis

- Cosine similarities against attribute class change
  - Randomly sample 100 latent codes, and then edit the images
  - Compute the cosine similarities with the initial direction



Failure Case Discussion

- Identity loss
  - Dataset bias and mode collapse issue of pretrained GAN
  - a small number of females with eyeglasses
- Artifacts
  - Many update iterations on latent codes would make the latent code fall into outlier region of the latent space
- Real Cases
  - GAN-inversion, an ill-posed problem
  - Introduce an additional gap between inverted latent code and the original latent code


(a) Identity Loss


(b) Artifacts

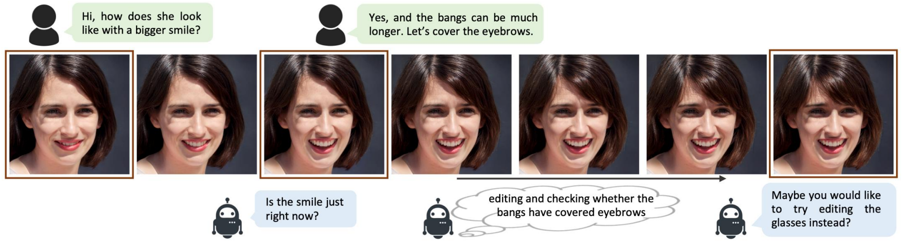

(c) Real Cases

Summary

Task

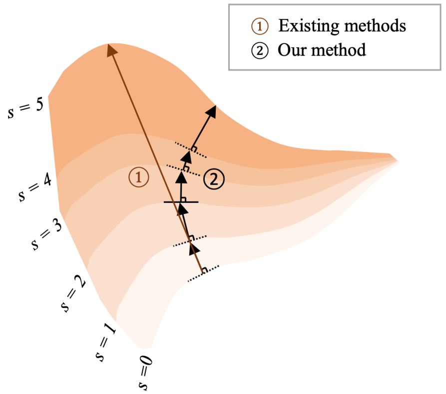
Dialog-based

Fine-Grained Facial Editing



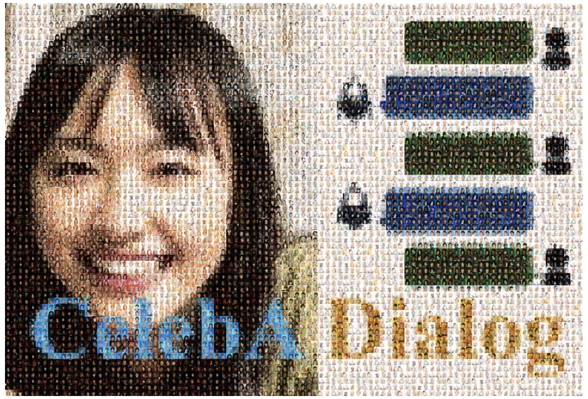
Method

Semantic Field



Dataset

CelebA-Dialog



Code and Models

Code

CelebA-Dialog Dataset

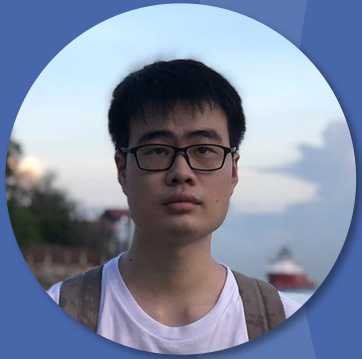
https://www.mmlab-ntu.com/

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

商汤 sensetime

THE PREMIER CONFERENCE & EXHIBITION ON
COMPUTER GRAPHICS & INTERACTIVE TECHNIQUES


Yuming Jiang[1]


Shuai Yang[1]


Haonan Qiu[1]


Wayne Wu[2]


Chen Change Loy[1]


Ziwei Liu[1]

[1]S-Lab Nanyang Technological University
[2]SenseTime Research

Text2Human

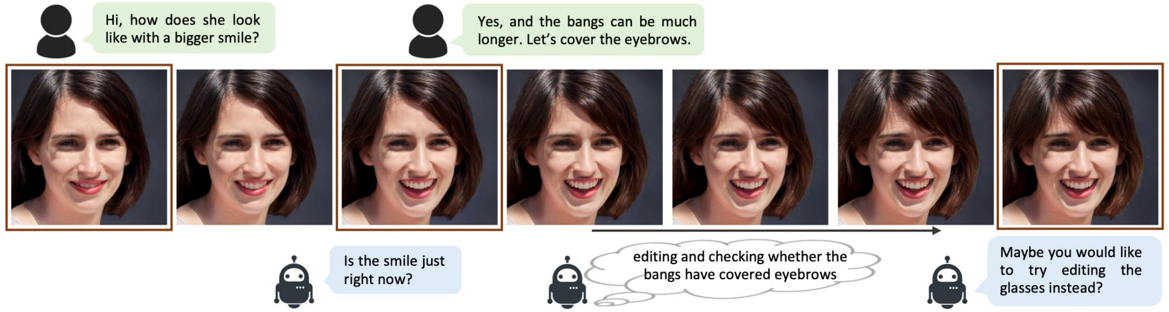
TEXT-DRIVEN CONTROLLABLE HUMAN IMAGE GENERATION

INTRODUCTION

- Generative Adversarial Networks

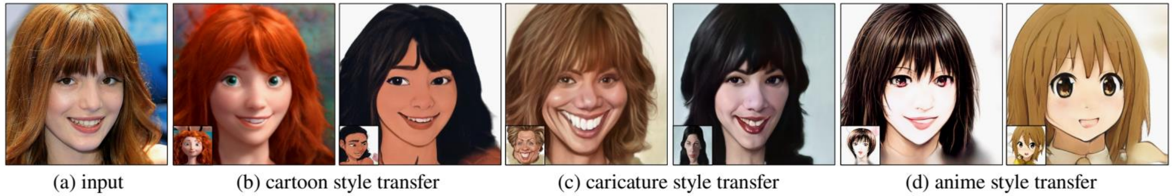
StyleGAN [Karras et al. 2018, 2020]

- Facial attribute editing



Talk-to-Edit [Jiang et al. 2021]

- Face Stylization

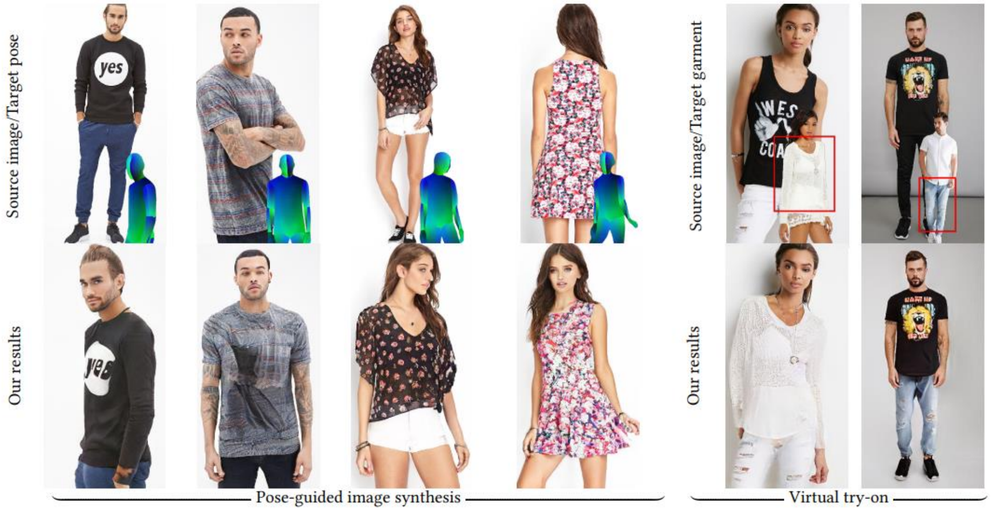


DualStyleGAN [Yang et al. 2022]

INTRODUCTION

- Human full-body images

- Pose Transfer
- Virtual try-on

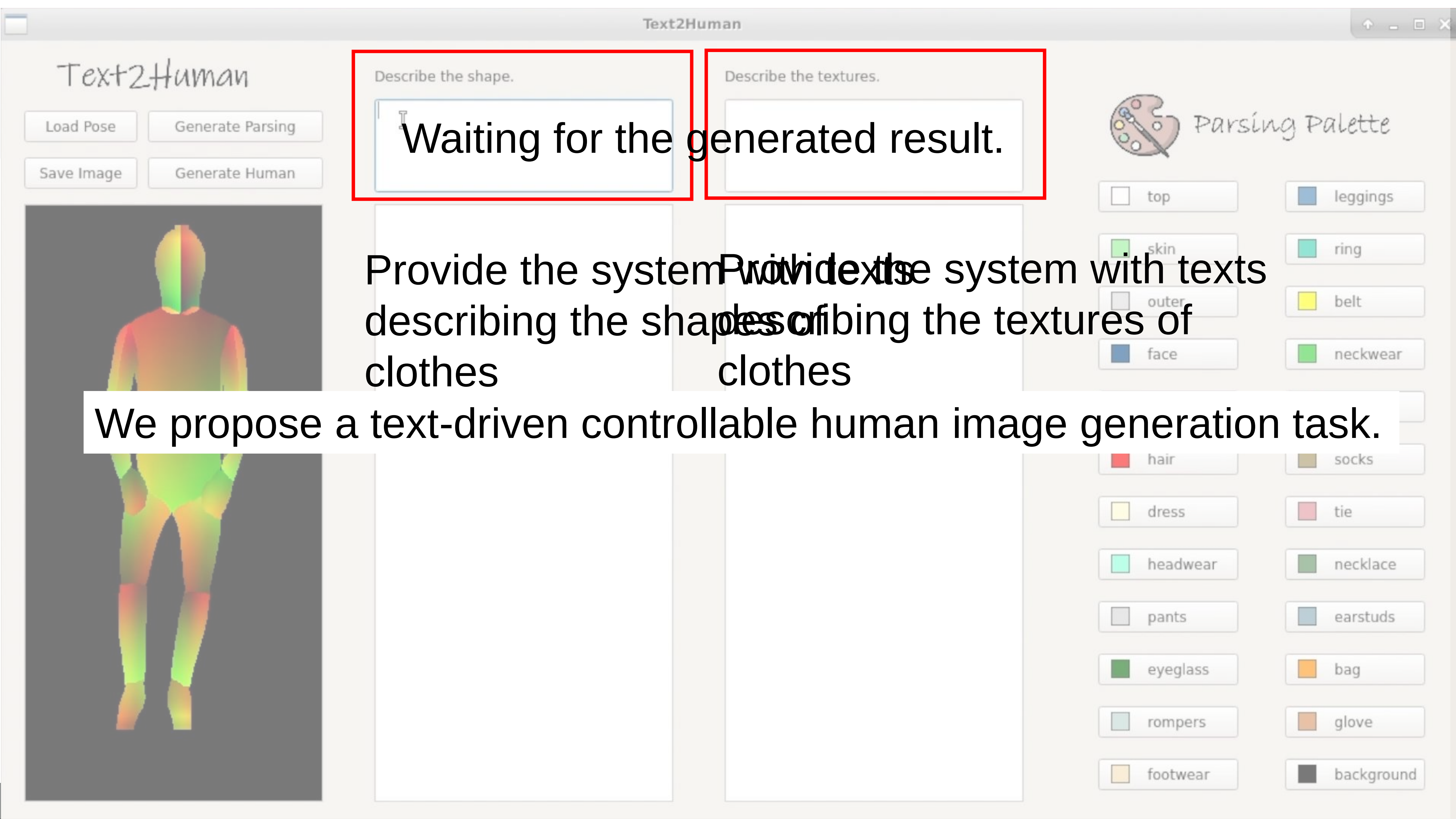


Pose with Style [Albahar et al. 2021]

INTRODUCTION

- Controllable human body image generation
  - More complex with multiple factors
  - Diverse styles of clothes
  - Textual controls need fine-grained annotations

PIPELINE OVERVIEW



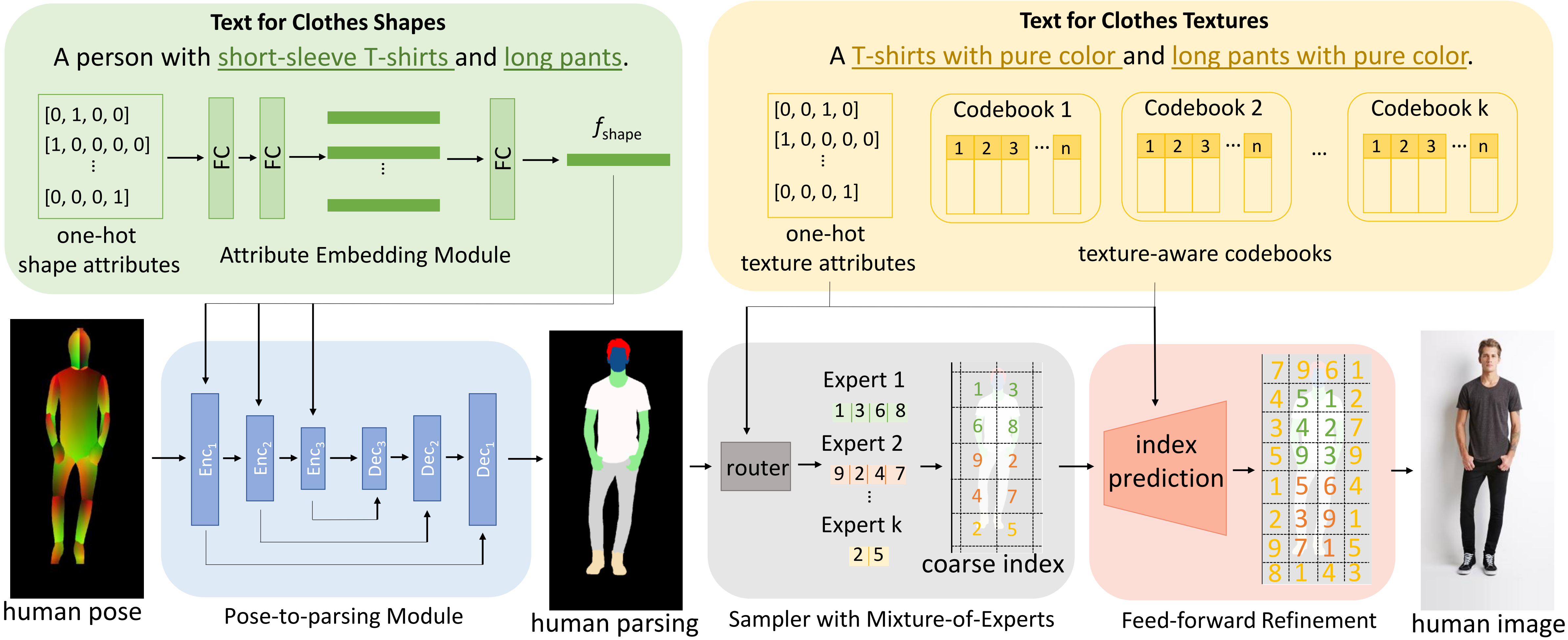
Waiting for the generated result.

Provide the system with texts describing the shapes of clothes

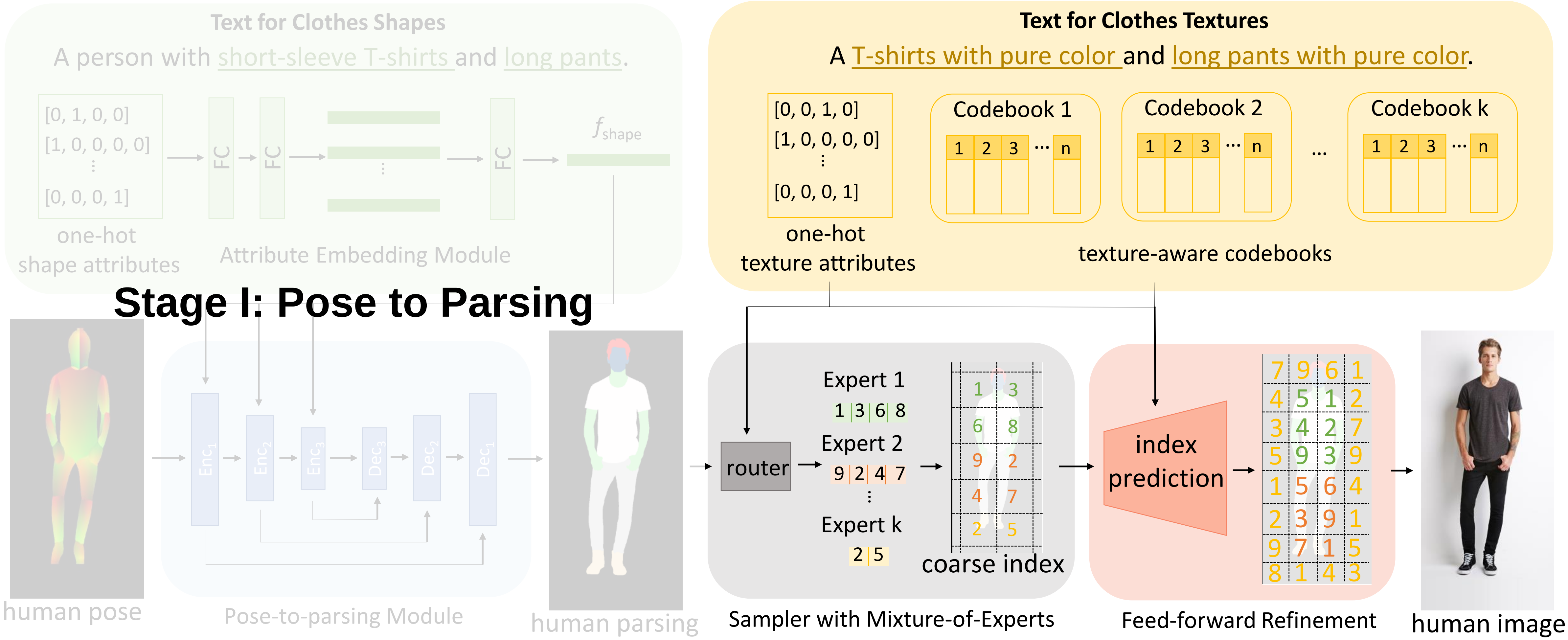
Provide the system with texts describing the textures of clothes

We propose a text-driven controllable human image generation task.

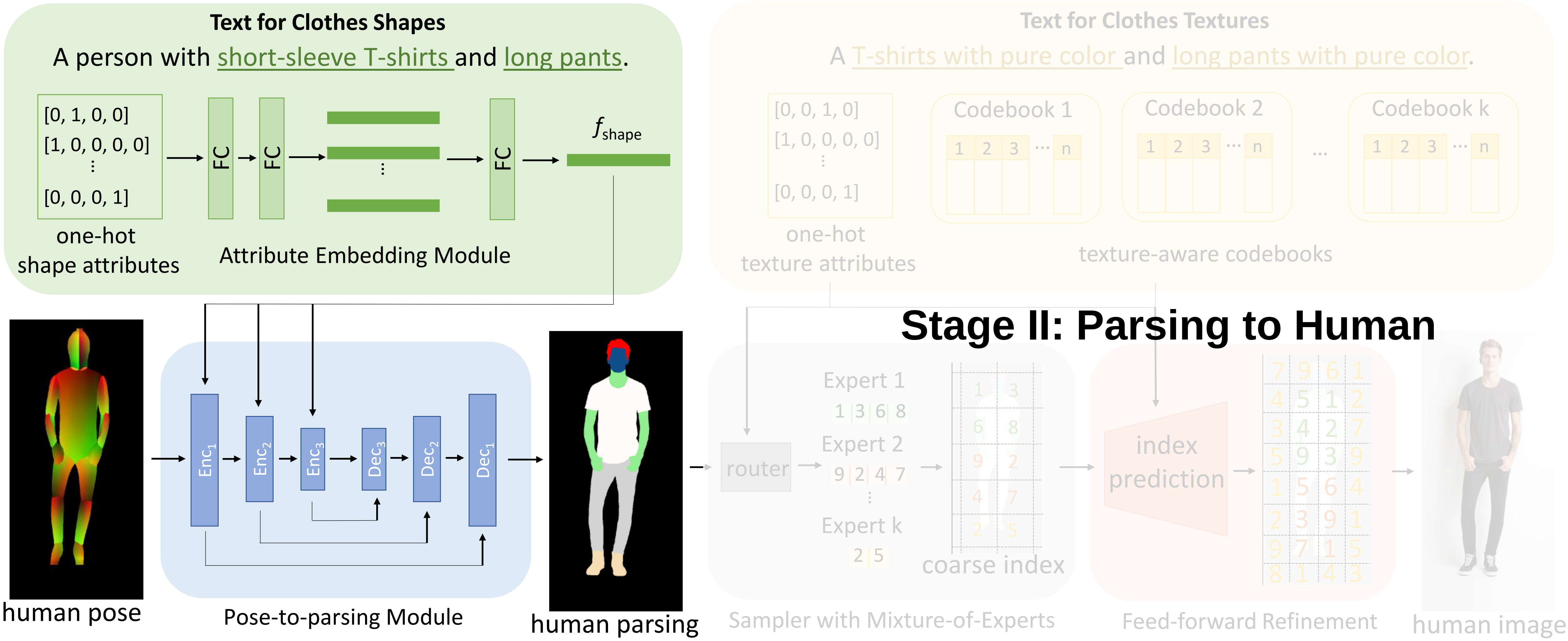
FRAMEWORK OF TEXT2HUMAN



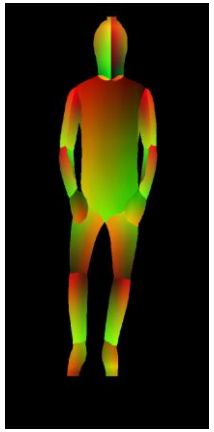
FRAMEWORK OF TEXT2HUMAN



FRAMEWORK OF TEXT2HUMAN



FRAMEWORK OF TEXT2HUMAN



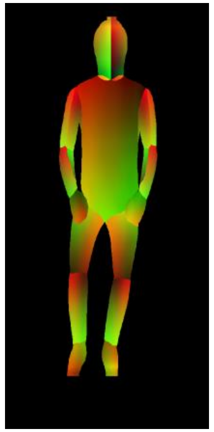
human pose

FRAMEWORK OF TEXT2HUMAN

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

Text for Clothes Shapes

A person with short-sleeve T-shirts and long pants.



human pose

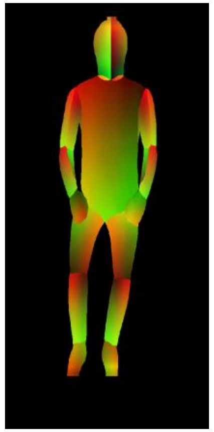
FRAMEWORK OF TEXT2HUMAN

Text for Clothes Shapes

A person with short-sleeve T-shirts and long pants.

[0, 1, 0, 0]
[1, 0, 0, 0, 0]
⋮
[0, 0, 0, 1]

one-hot
shape attributes



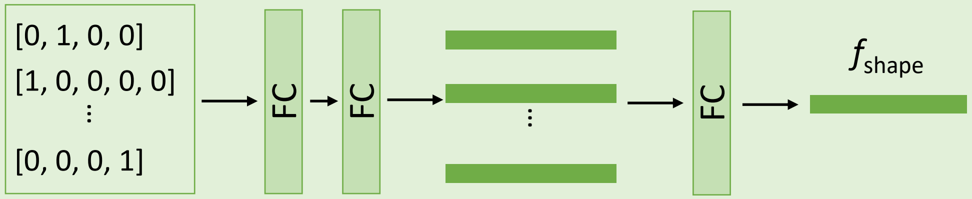
human pose

FRAMEWORK OF TEXT2HUMAN

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

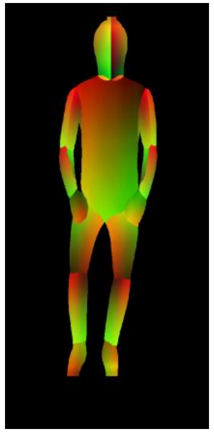
Text for Clothes Shapes

A person with short-sleeve T-shirts and long pants.



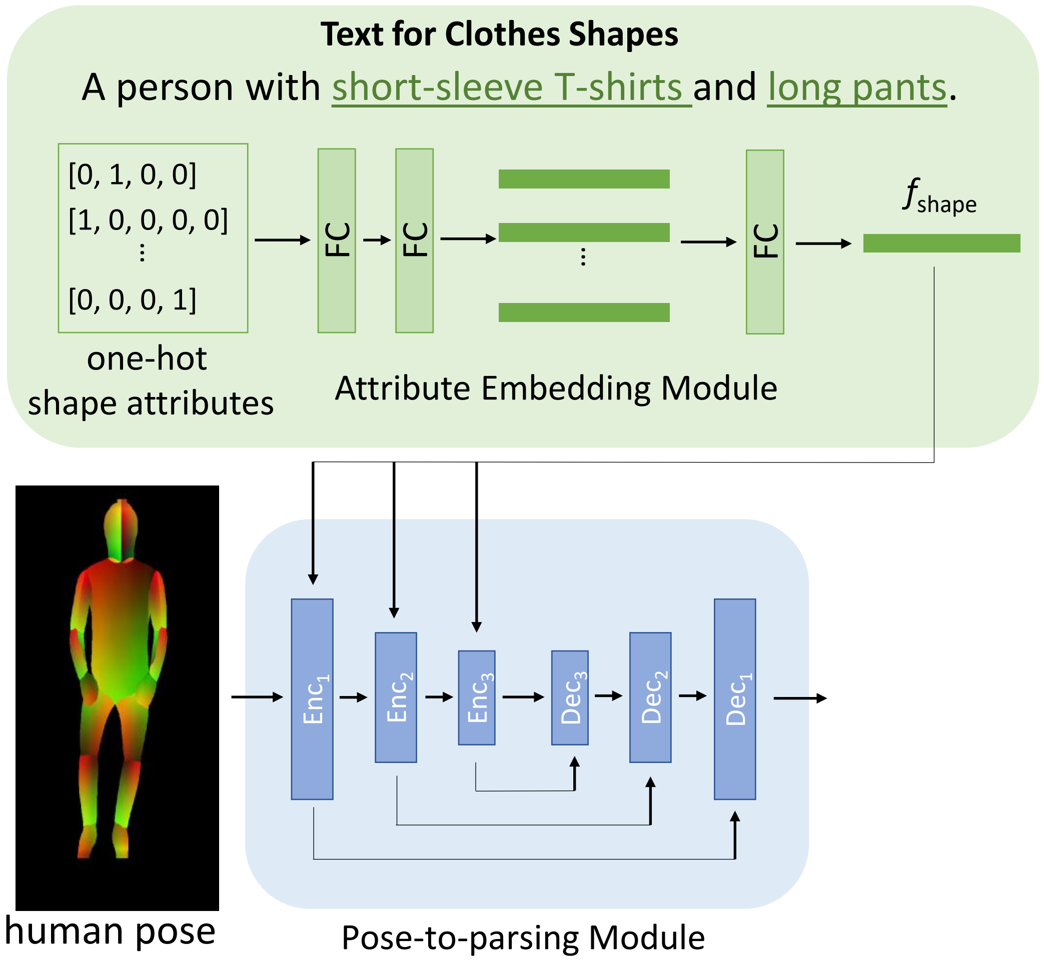
one-hot shape attributes

Attribute Embedding Module



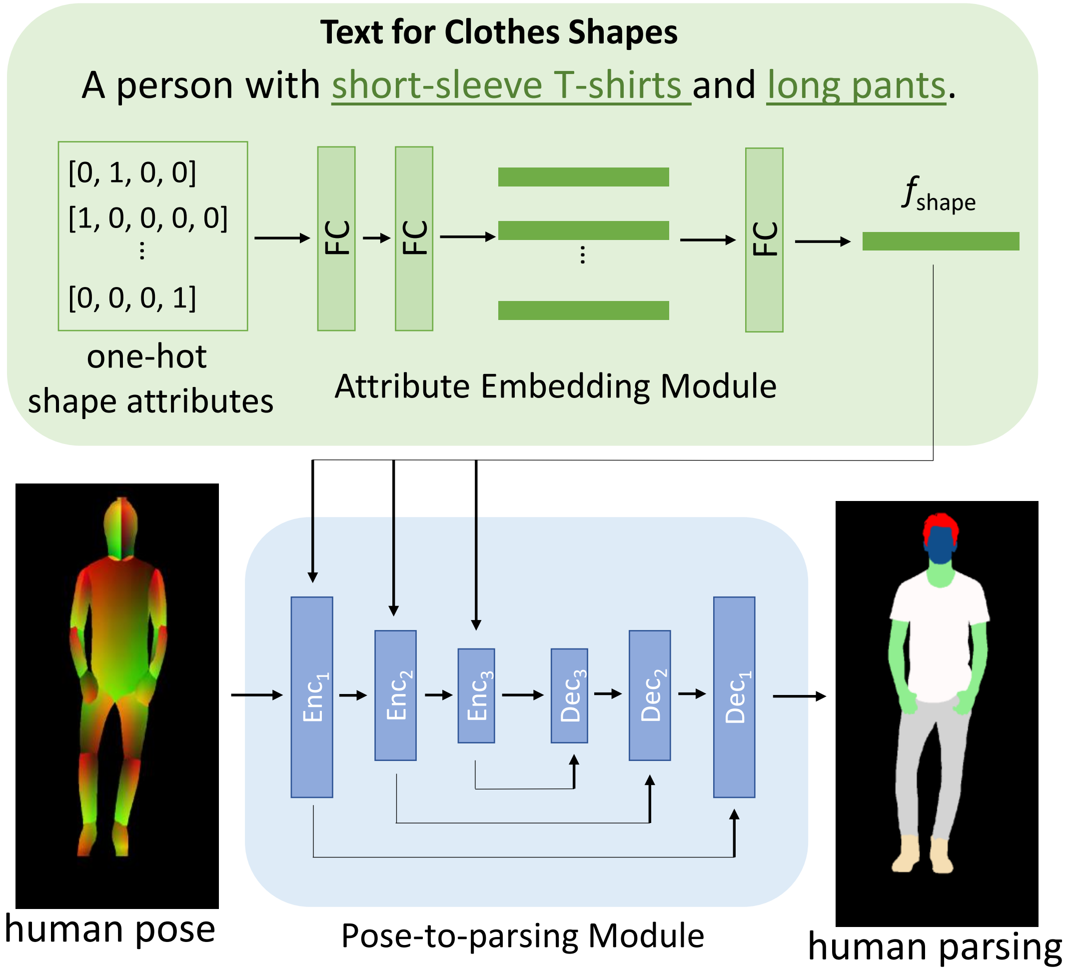
human pose

FRAMEWORK OF TEXT2HUMAN

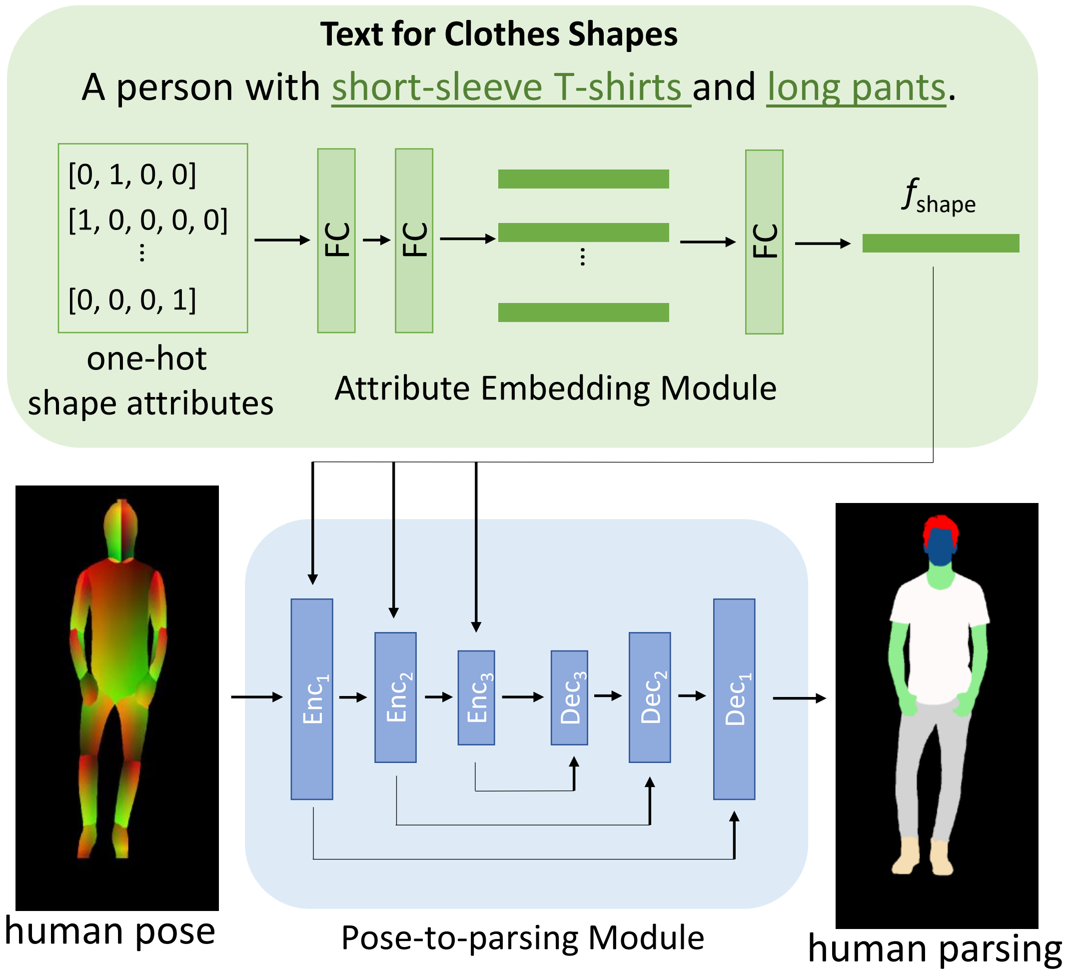


FRAMEWORK OF TEXT2HUMAN

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG



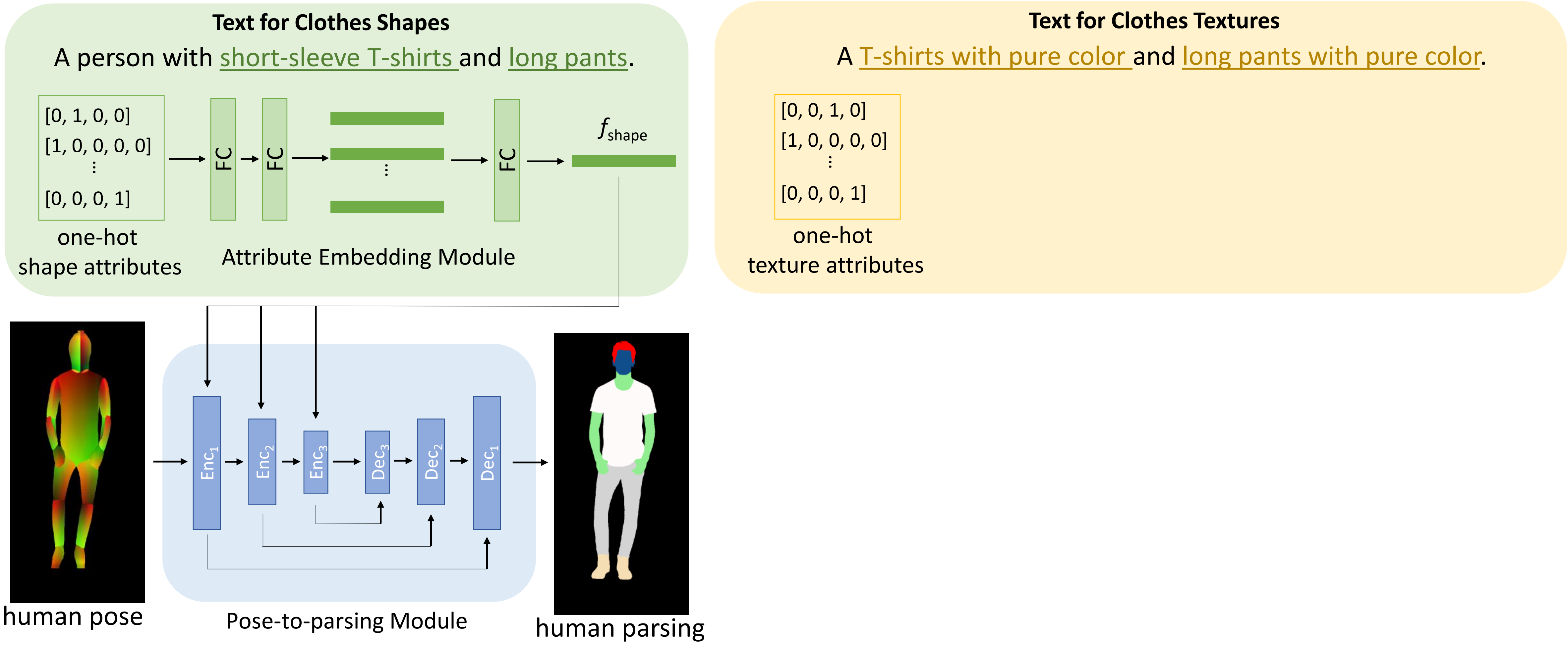
FRAMEWORK OF TEXT2HUMAN



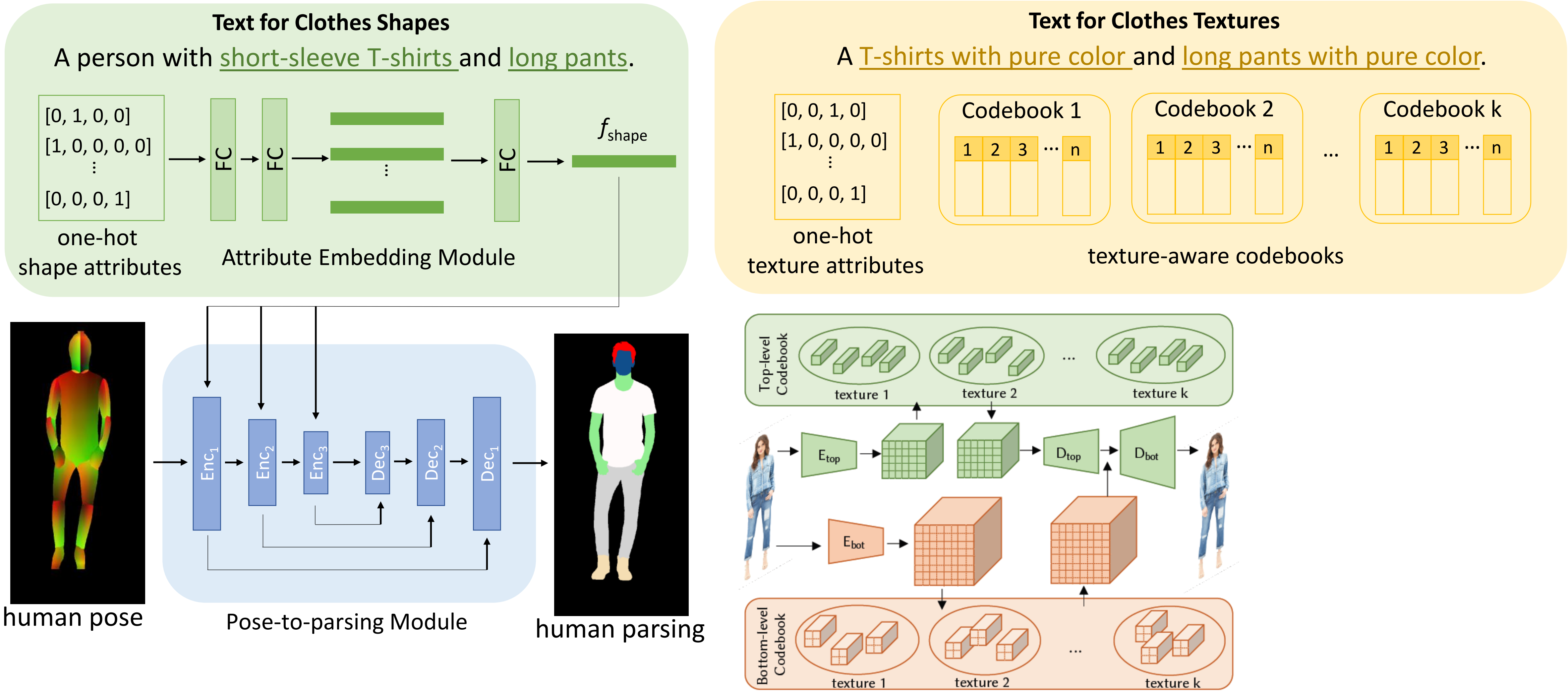
Text for Clothes Textures

A T-shirts with pure color and long pants with pure color.

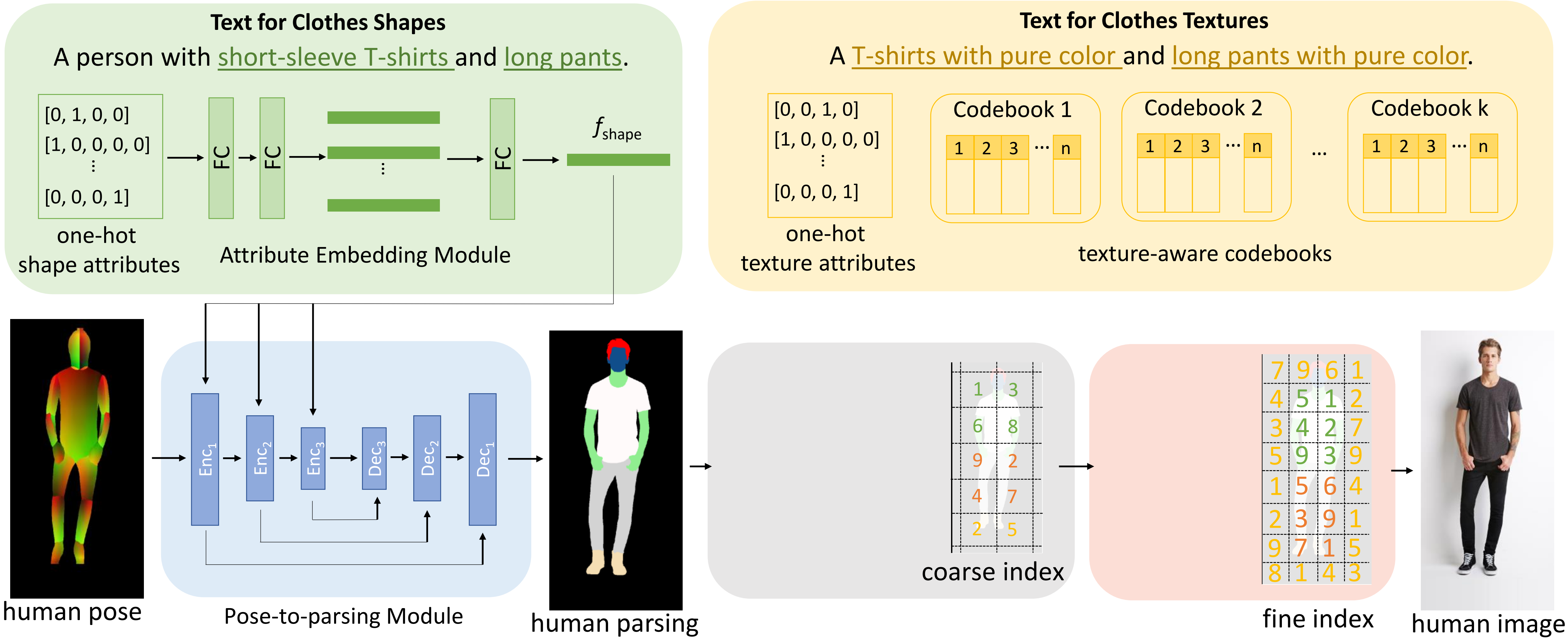
FRAMEWORK OF TEXT2HUMAN



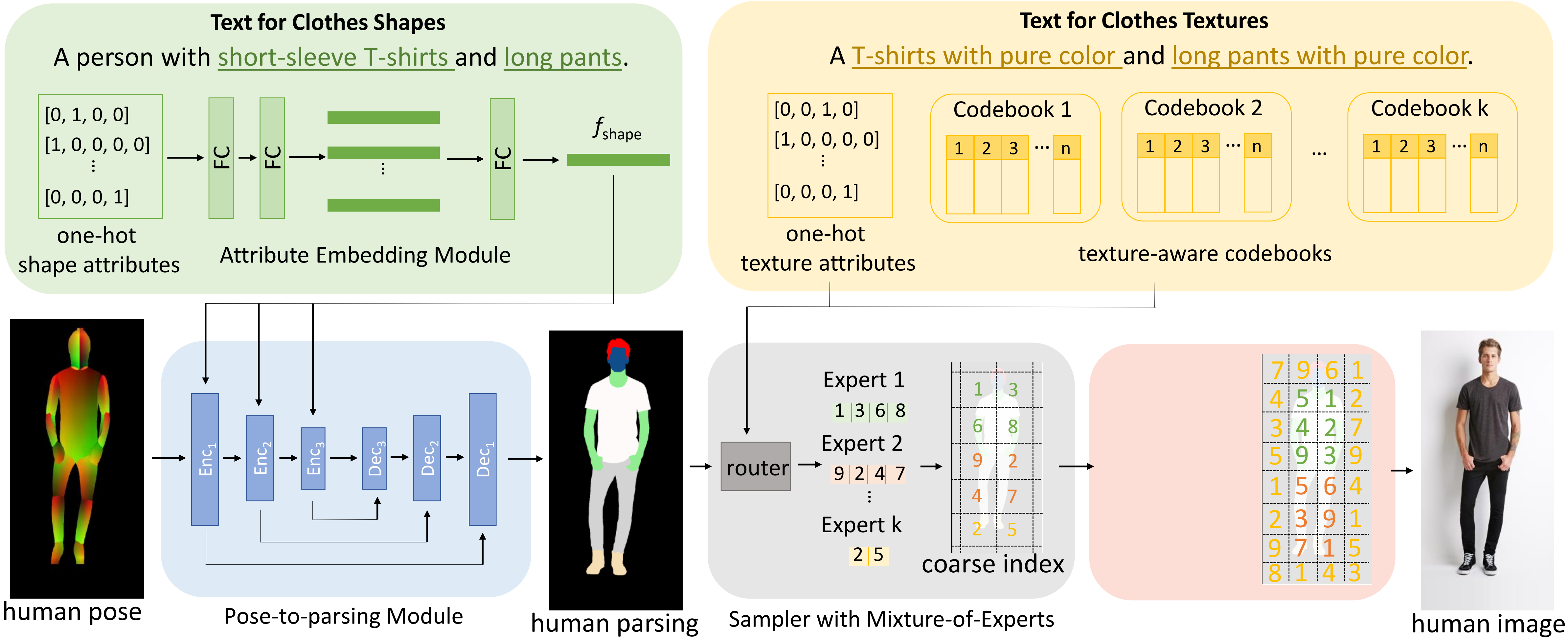
FRAMEWORK OF TEXT2HUMAN



FRAMEWORK OF TEXT2HUMAN

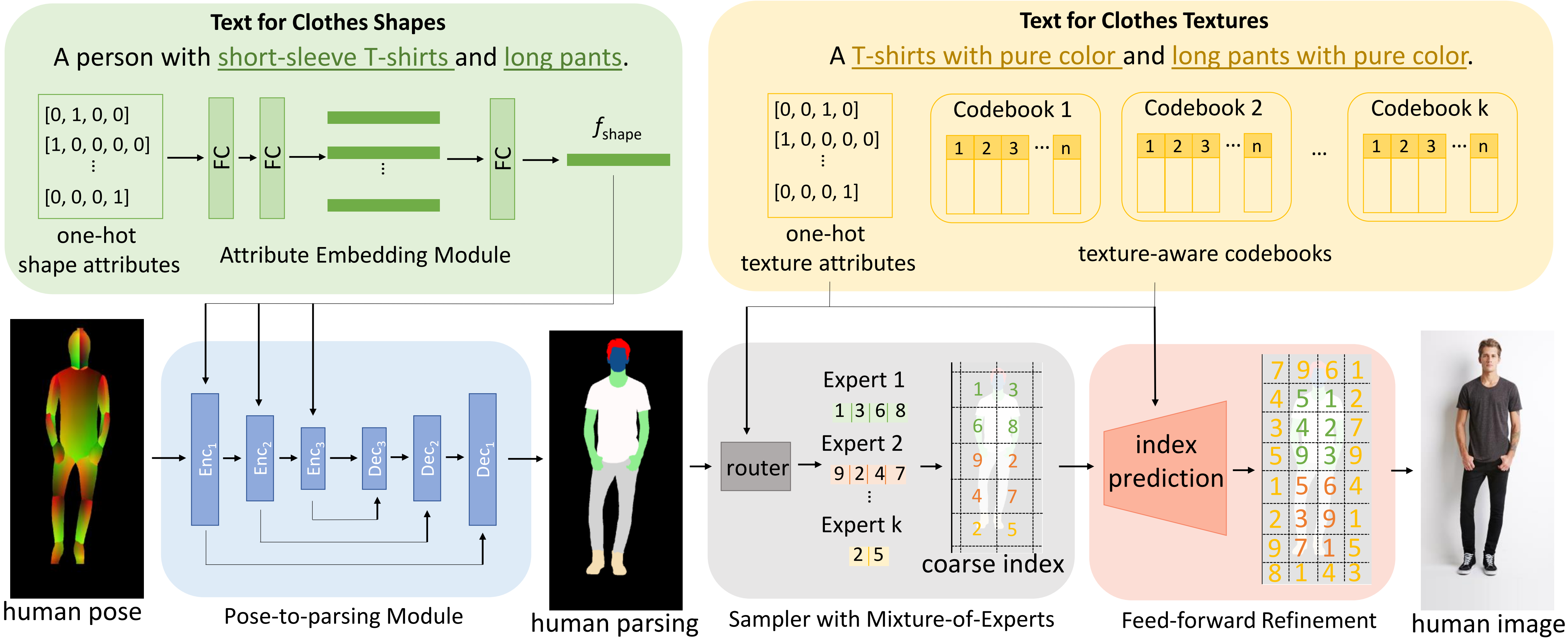


FRAMEWORK OF TEXT2HUMAN



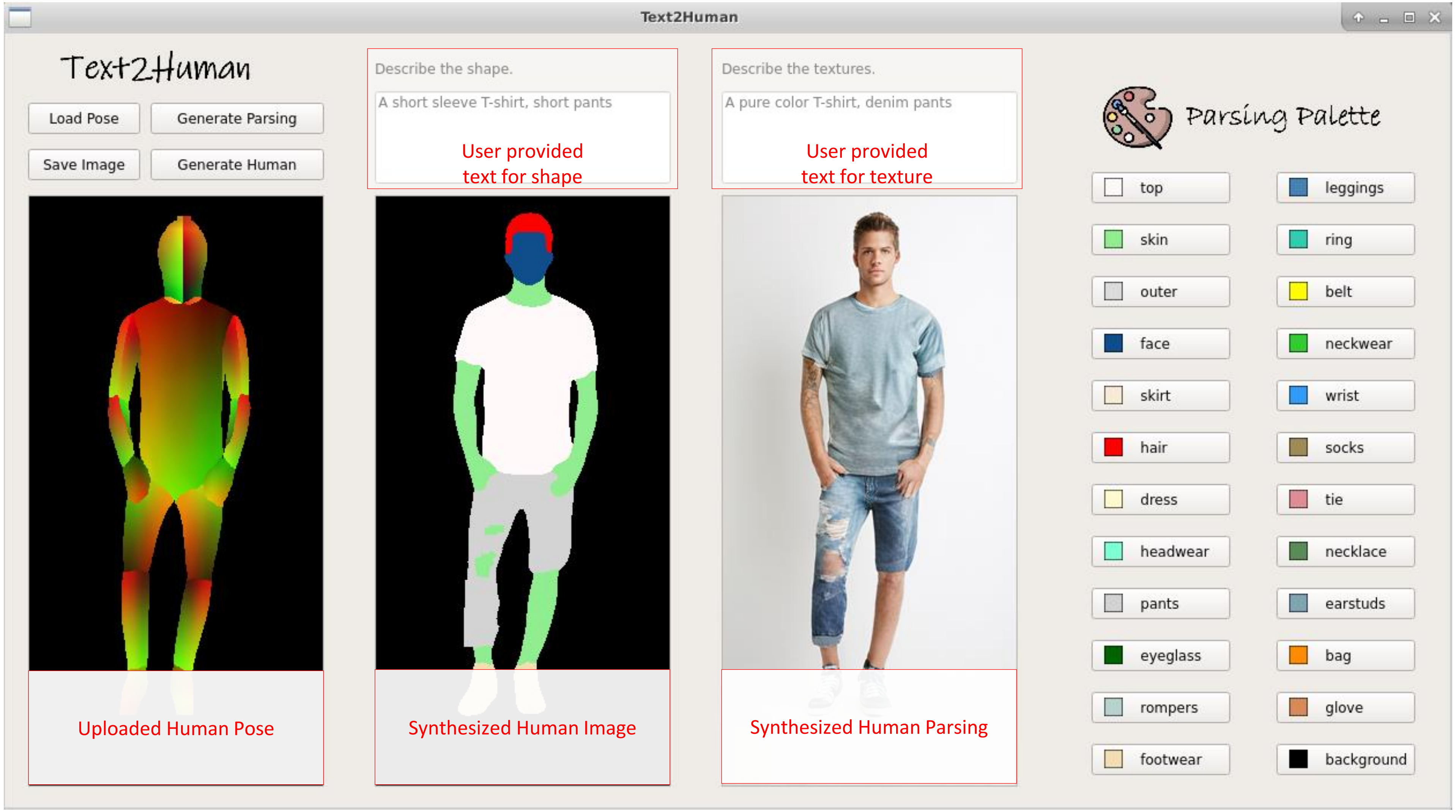
FRAMEWORK OF TEXT2HUMAN

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG



INTERACTIVE USER INTERFACE

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG



DEEPFASHION-MULTIMODAL DATASET

DEEPFASHION-MULTIMODAL DATASET

- 44,096 high-resolution human images, including 12,701 full body human images
- **manually annotated** the human parsing labels
- DensePose for each human image
- **manually annotated** the keypoints
- **manually annotated** with attributes
- textual description

human image | human parsing | densepose | key points | labels | textual descriptions

EXPERIMENT


pure color upper clothes with a denim outer, seven-point and pure color pants
floral upper clothes with a pure color outer, three-point jeans
Parsing
Pix2PixHD
SPADE
MISC
HumanGAN
Text2Human

EXPERIMENT


Parsing


Taming Transformer


Text2Human


Parsing

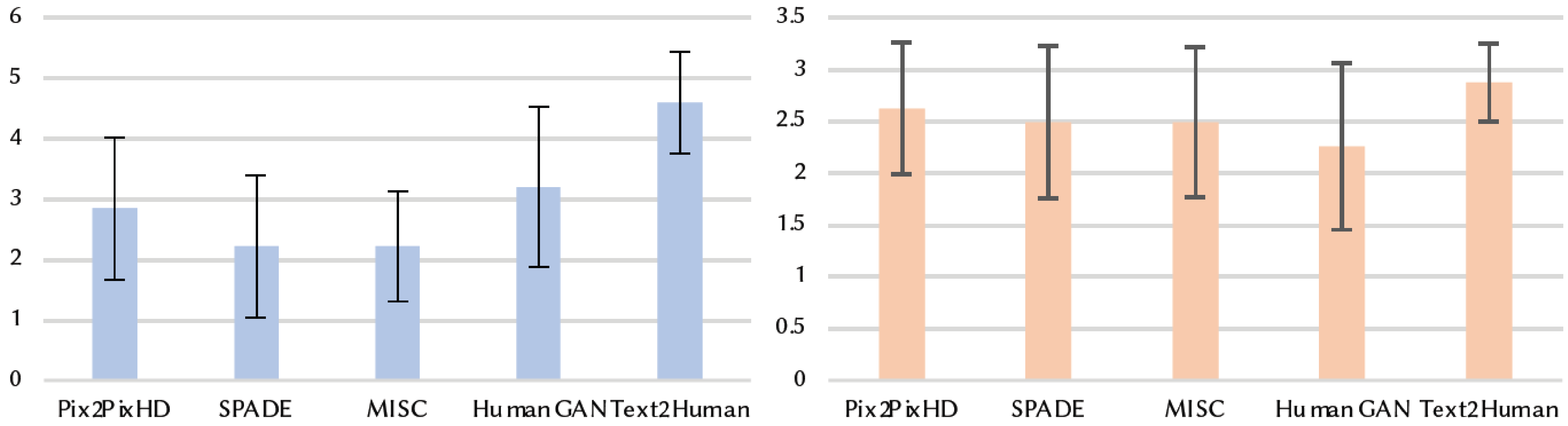

Taming Transformer


Text2Human

EXPERIMENT

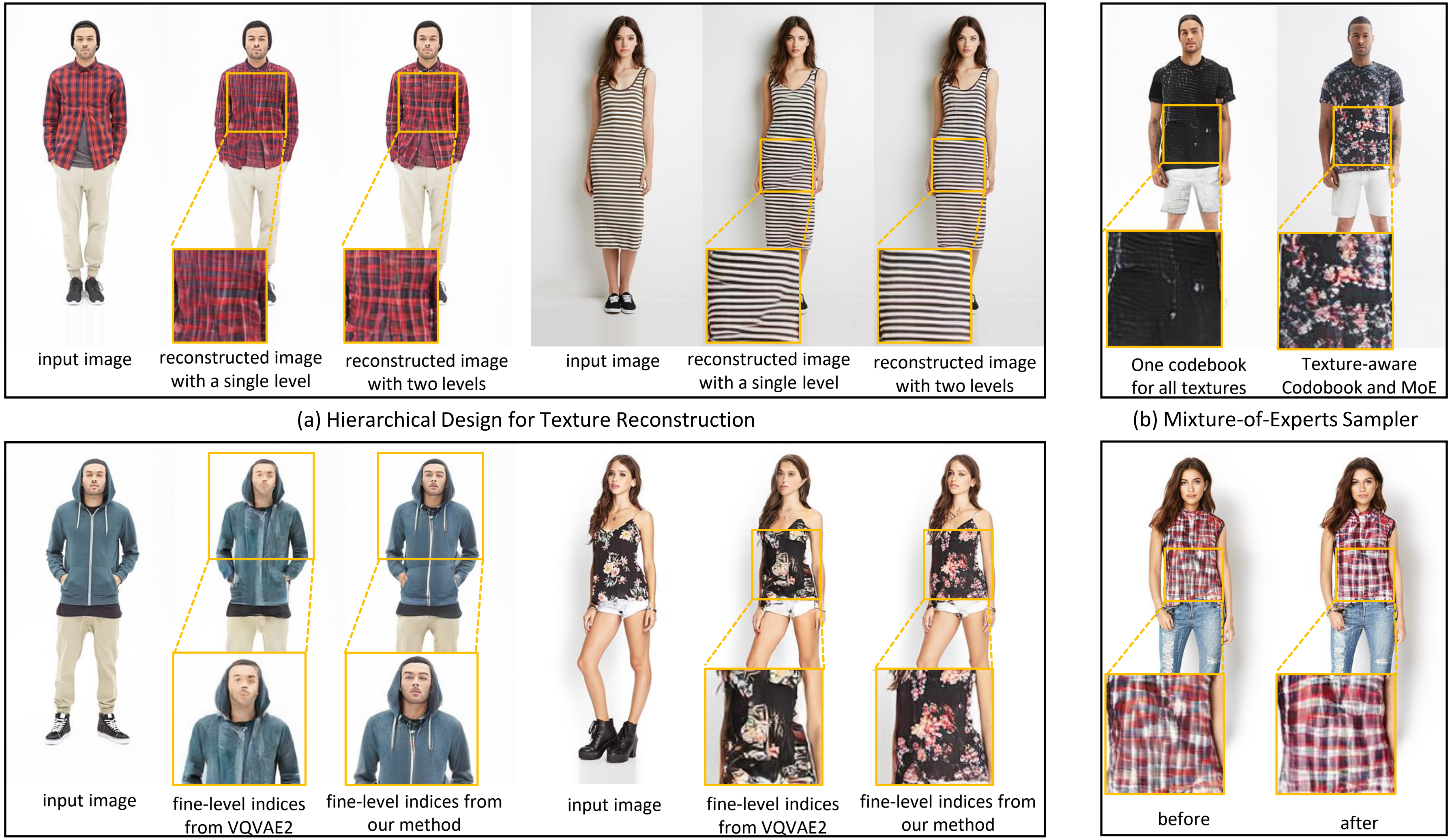
Pose TryOnGAN HumanGAN Text2Human

EXPERIMENT



(a) photorealism (b) texture consistency score

ABLATION STUDY



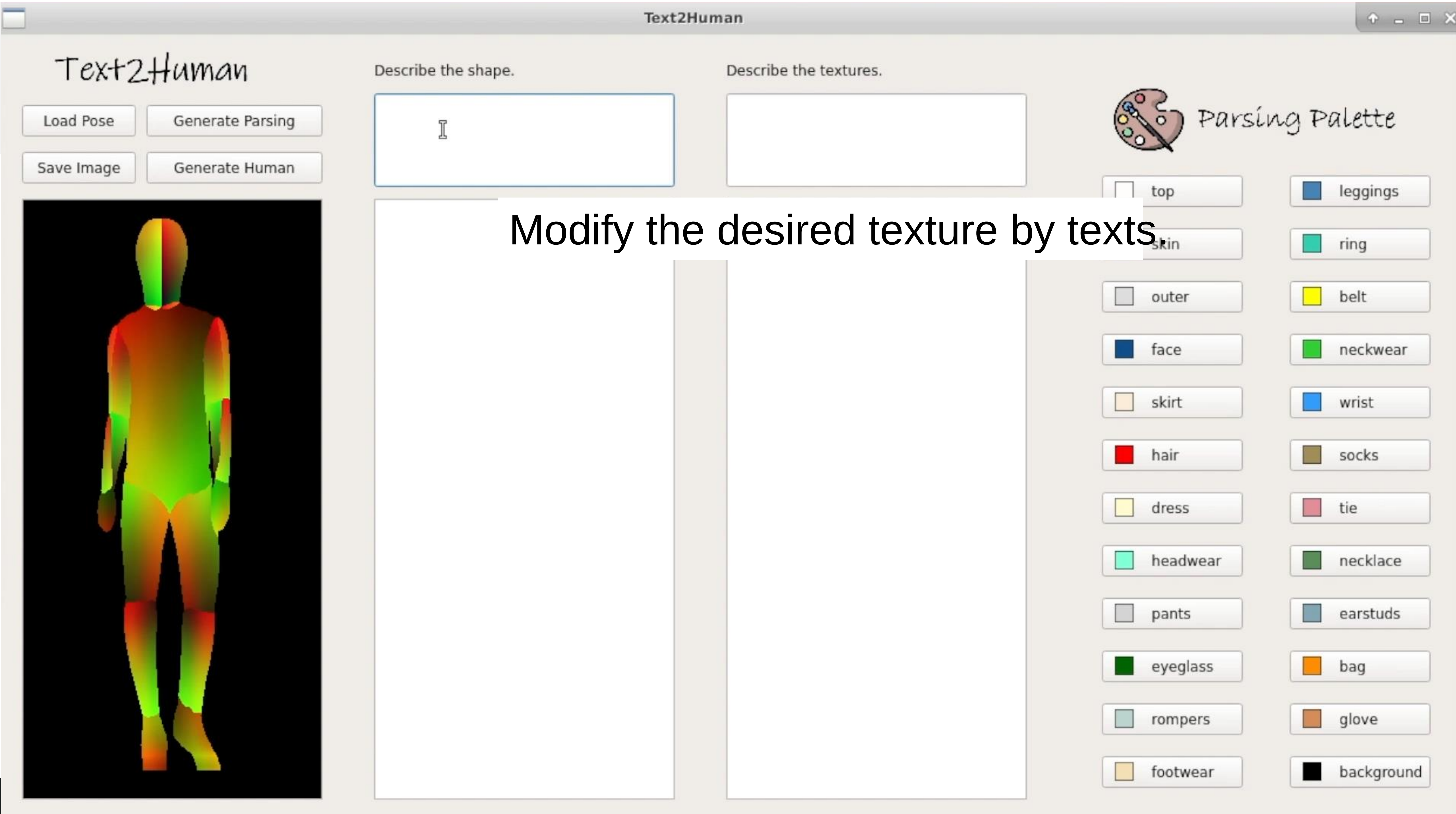
(a) Hierarchical Design for Texture Reconstruction

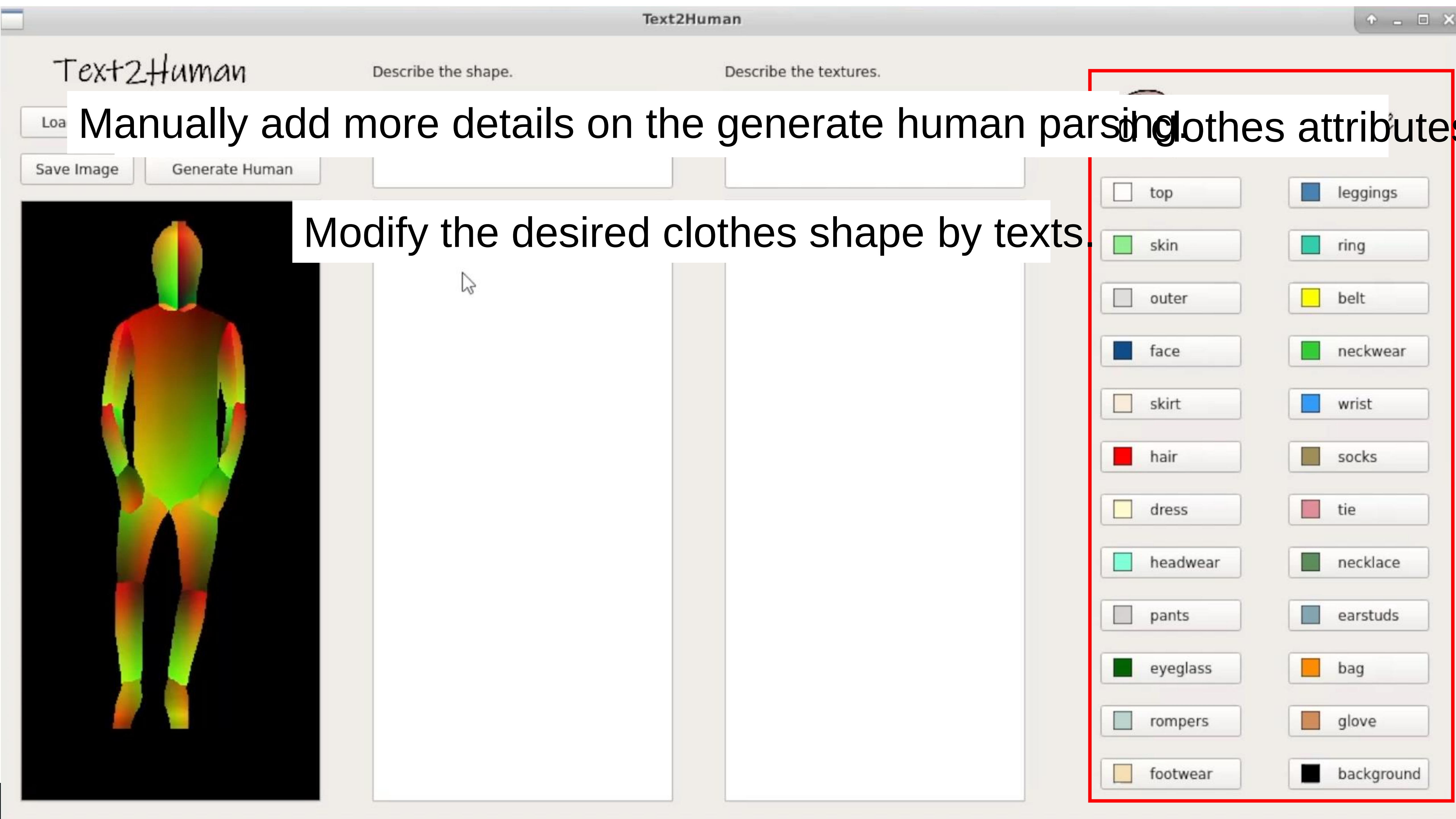
(b) Mixture-of-Experts Sampler

(c) Effectiveness of Feed-forward Index Prediction Network

(d) Refinement

MORE INTERACTIVE EXAMPLES


Text2Human
Text2Human
Describe the shape.
Describe the textures.
Load Pose
Generate Parsing
Save Image
Generate Human
Parsing Palette
Modify the desired texture by texts.
top
leggings
skin
ring
outer
belt
face
neckwear
skirt
wrist
hair
socks
dress
tie
headwear
necklace
pants
earstuds
eyeglass
bag
rompers
glove
footwear
background

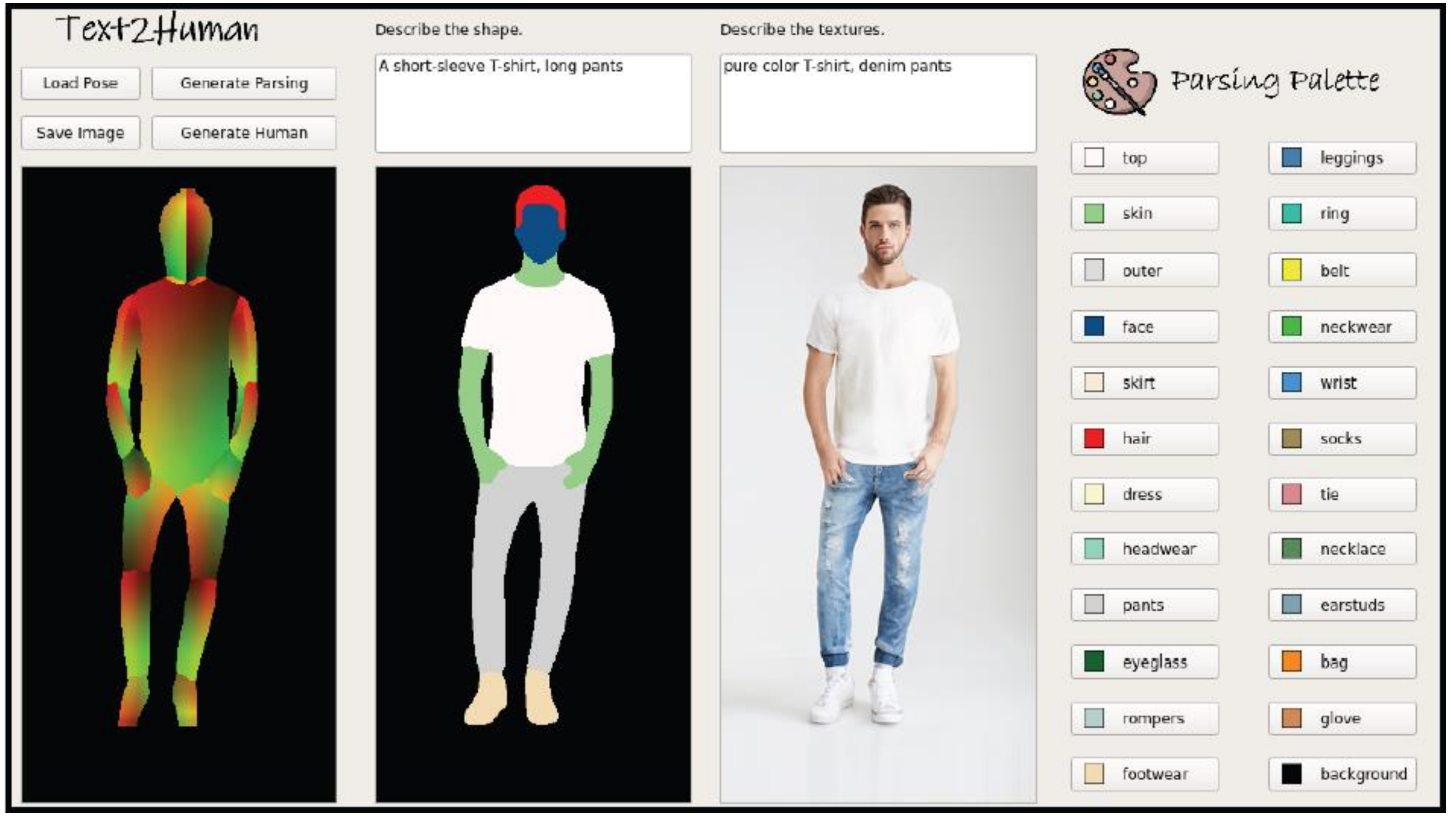

Text2Human
Text2Human
Describe the shape.
Describe the textures.
Manually add more details on the generate human parsing.
clothes attribute
Save Image
Generate Human
Modify the desired clothes shape by texts.
top
leggings
skin
ring
outer
belt
face
neckwear
skirt
wrist
hair
socks
dress
tie
headwear
necklace
pants
earstuds
eyeglass
bag
rompers
glove
footwear
background

MORE SYNTHESIZED HUMAN IMAGES

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

Task

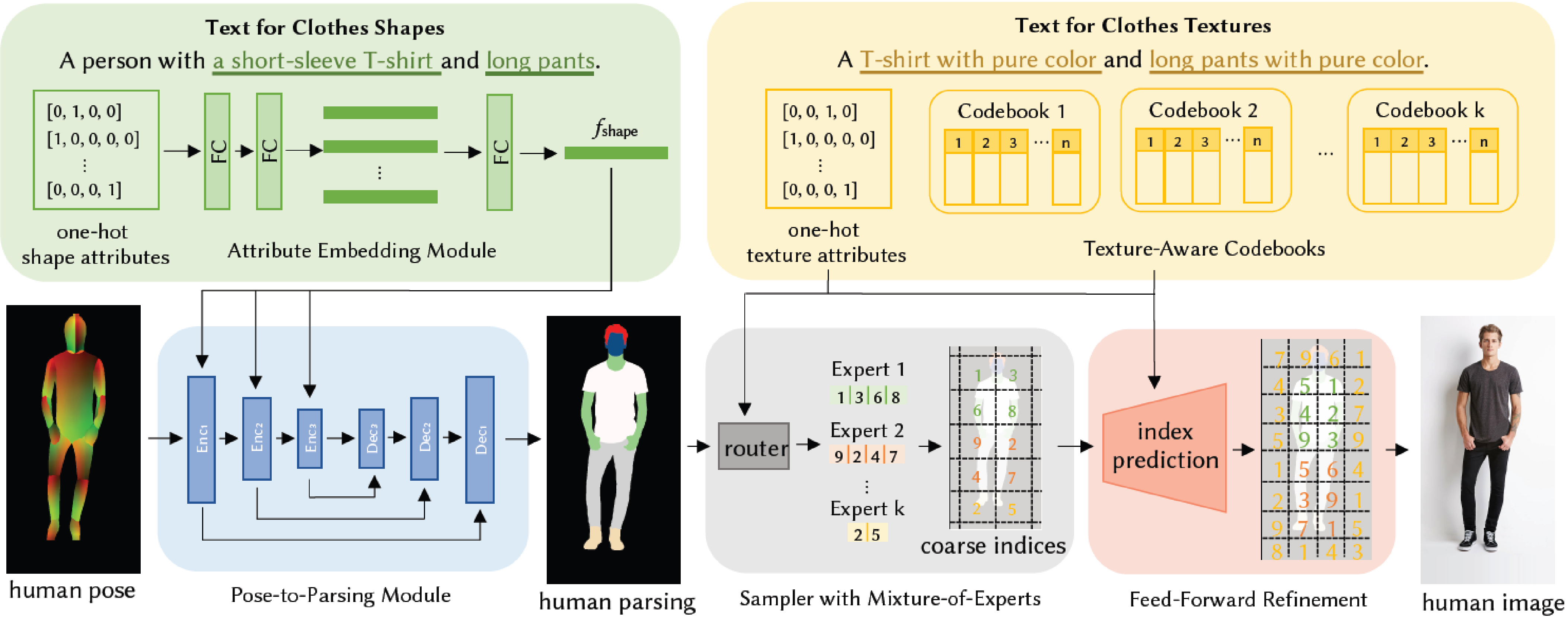
Controllable Human Image Generation



SIGGRAPH 2022 VANCOUVER+ 8-11 AUG

Method

Text2Human



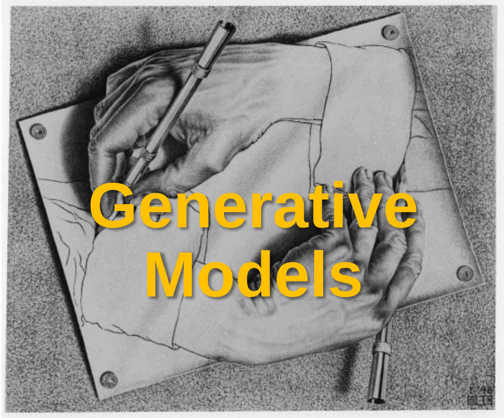
Dataset

DeepFashion-Multimodal

CODE AND MODELS

Thank You!

Human-Centric



Scaling

Interactive