AI-Driven Visual Content Generation

Ziwei Liu 刘子纬

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

MMLab@NTU

MMLab
@NTU

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

ABOUT
MMLab@NTU

MMLab@NTU was formed on the 1 August 2018, with a research focus on computer vision and deep learning. Its sister lab is MMLab@CUHK. It is now a group with three faculty members and more than 40 members including research fellows, research assistants, and PhD students.

Members in MMLab@NTU conduct research primarily in low-level vision, image and video understanding, creative content creation, 3D scene understanding and reconstruction. Have a look at the overview of our research. All publications are listed here.

We are always looking for motivated PhD students, postdocs, research assistants who have the same interests like us. Check out the careers page and follow us on Twitter.

AI-Generated Content (AIGC)

Movie

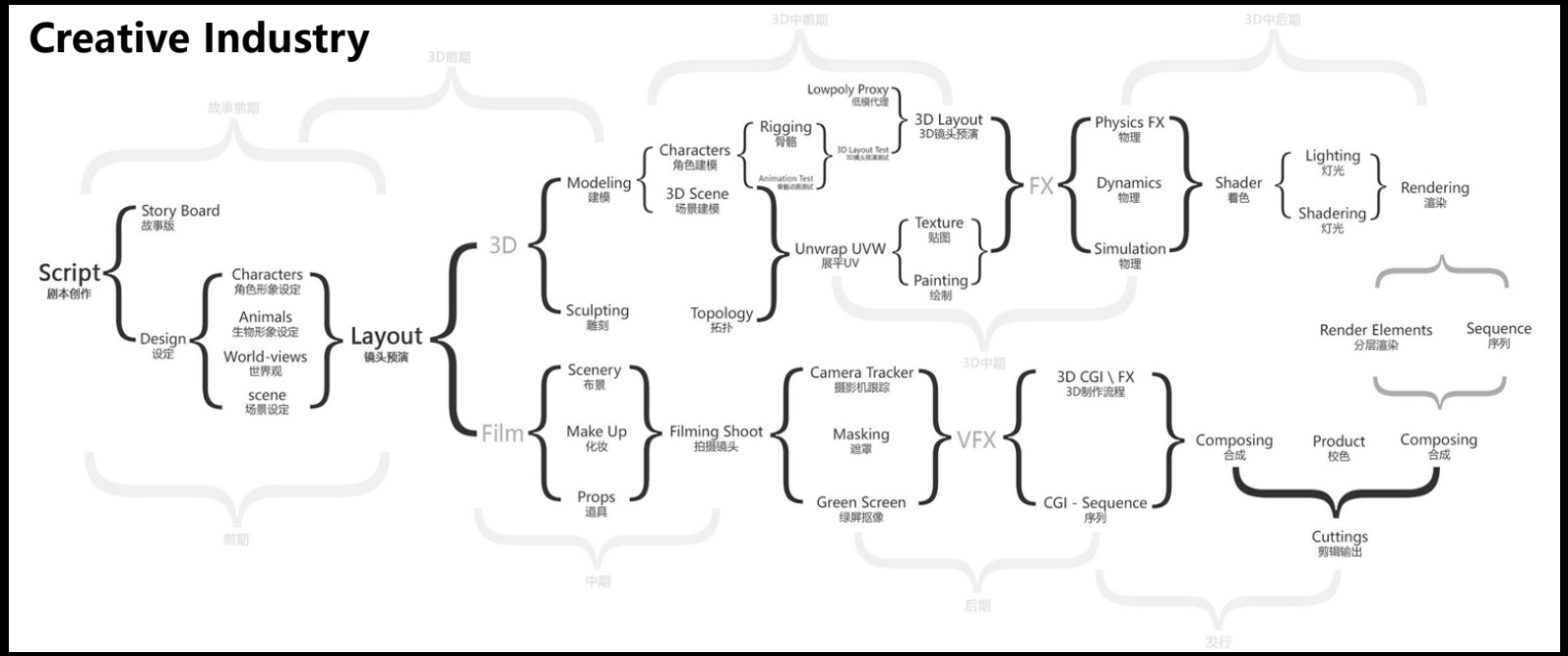
Game

Anime

VTuber

Virtual Beings

Creative Industry



2D Generation

3D Generation

Motion Generation

Scene Generation

2D Generation

3D Generation

Data

Software

Motion Generation

Scene Generation

2D Generation

3D Generation

Data

Software

Motion Generation

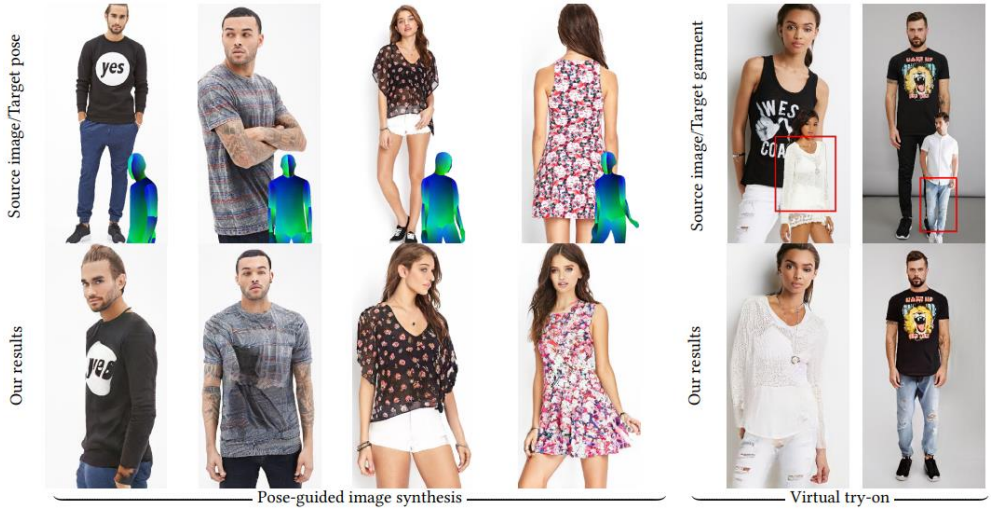
Scene Generation

INTRODUCTION

- Human full-body images



- Pose Transfer
- Virtual try-on



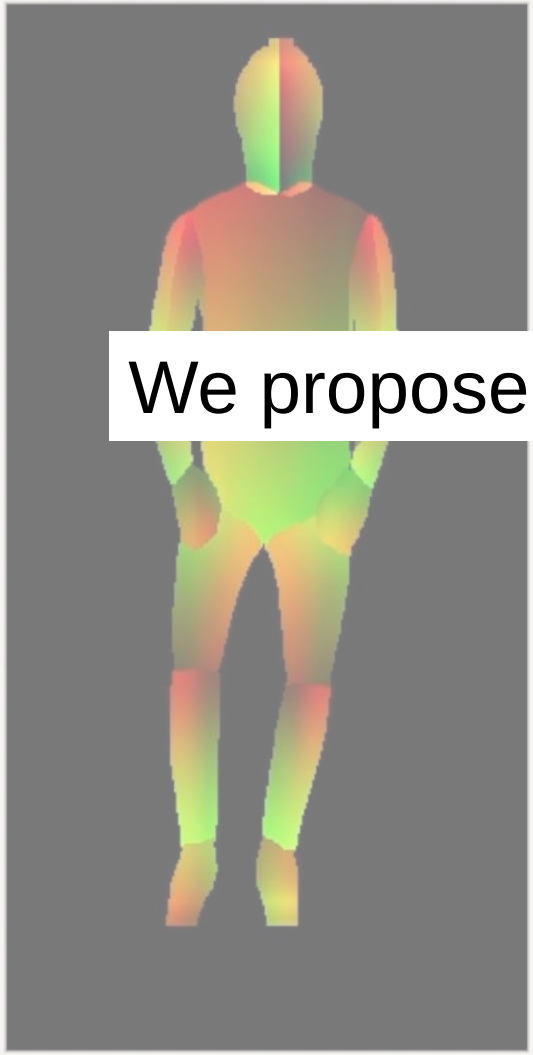
Pose with Style [Albahar et al. 2021]

Text2Human

Text2Human

Load Pose | Generate Parsing

Save Image | Generate Human



Describe the shape.

Waiting for the generated result.

Describe the textures.

Parsing Palette

top | leggings

skin | ring

outer | belt

face | neckwear

hair | socks

dress | tie

headwear | necklace

pants | earstuds

eyeglass | bag

rompers | glove

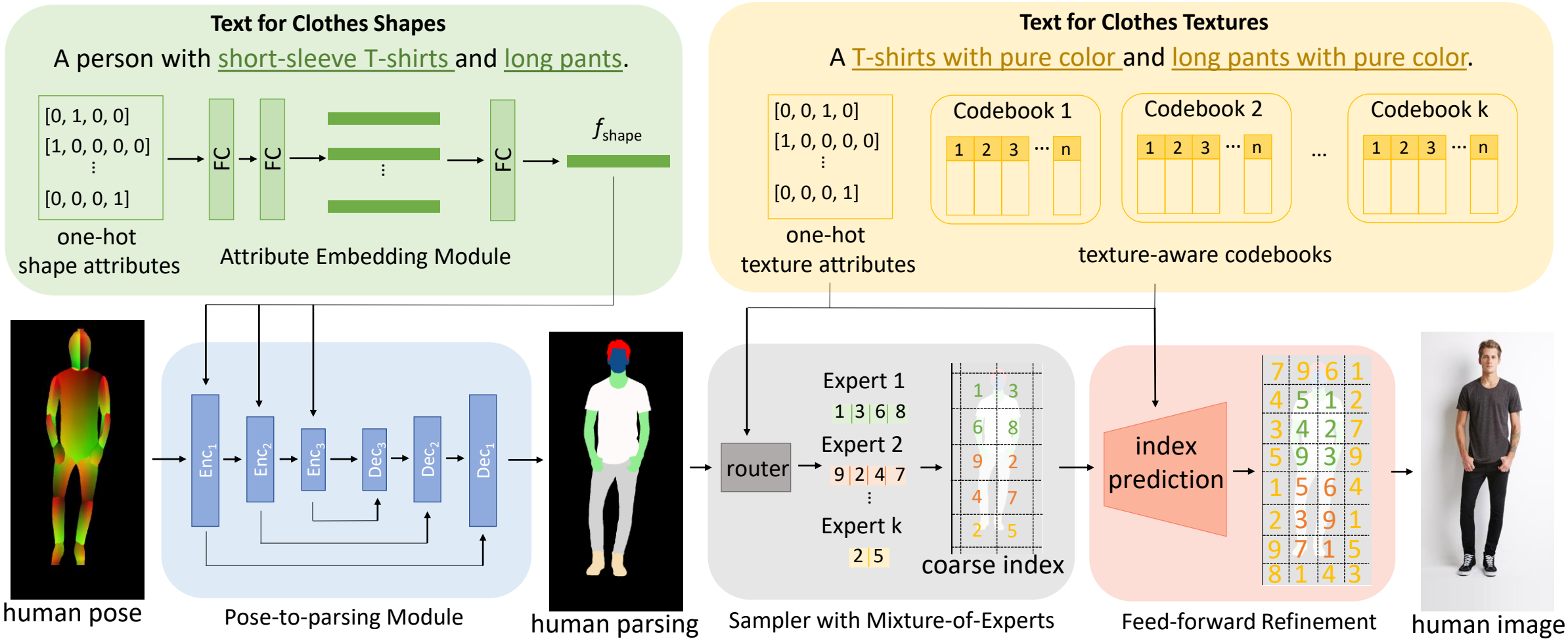
footwear | background

Provide the system with texts describing the shapes of clothes

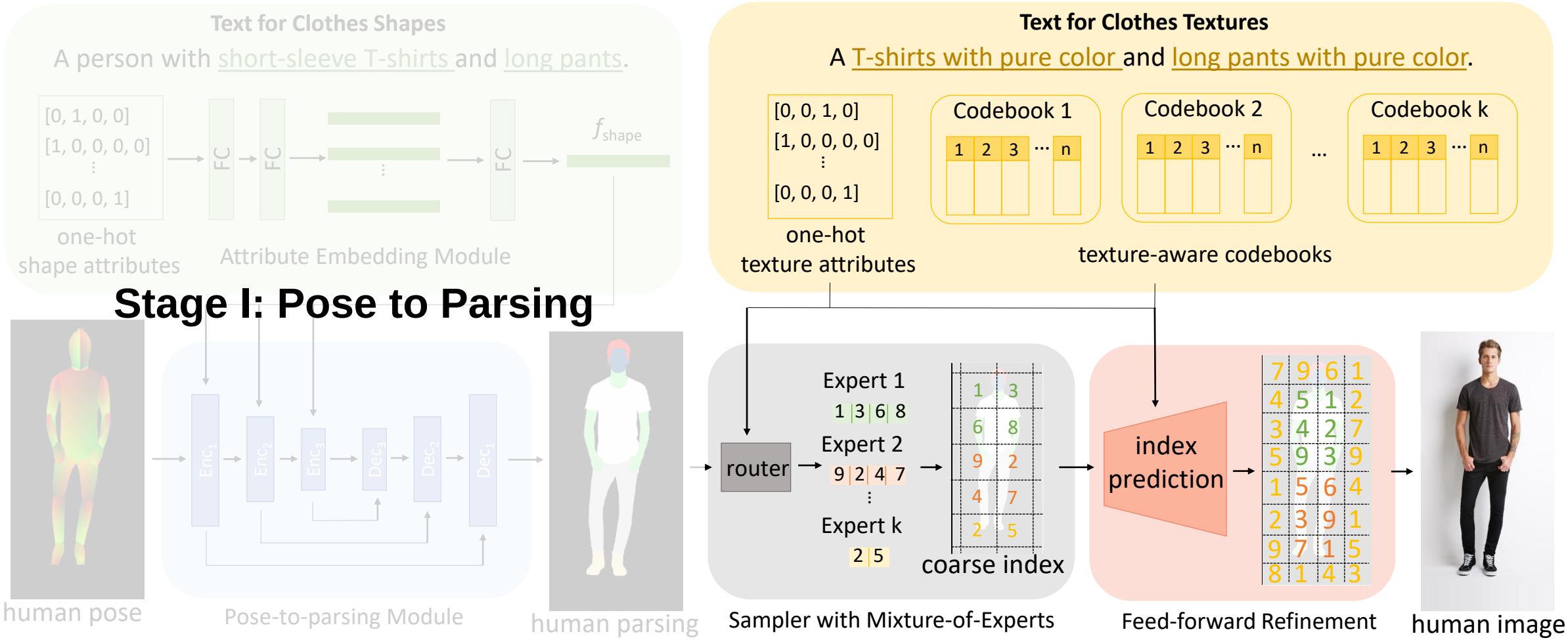
Provide the system with texts describing the textures of clothes

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

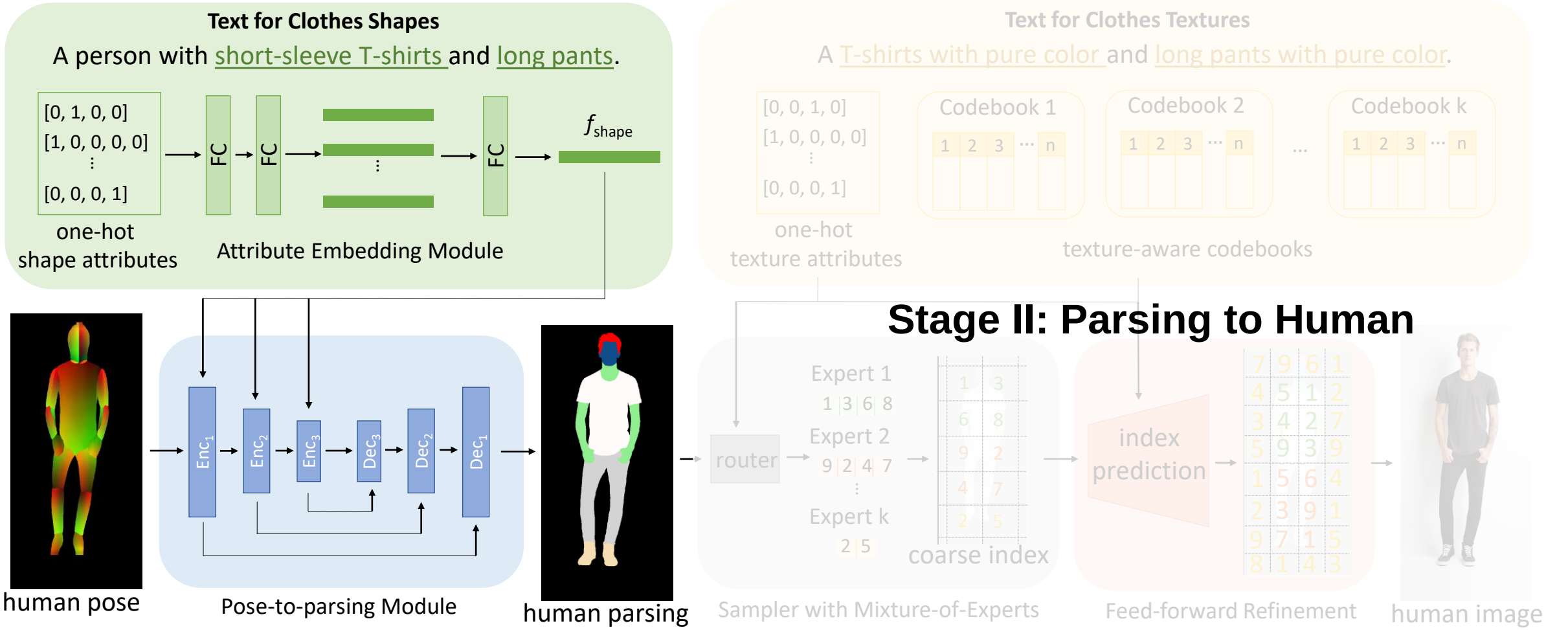
FRAMEWORK OF TEXT2HUMAN



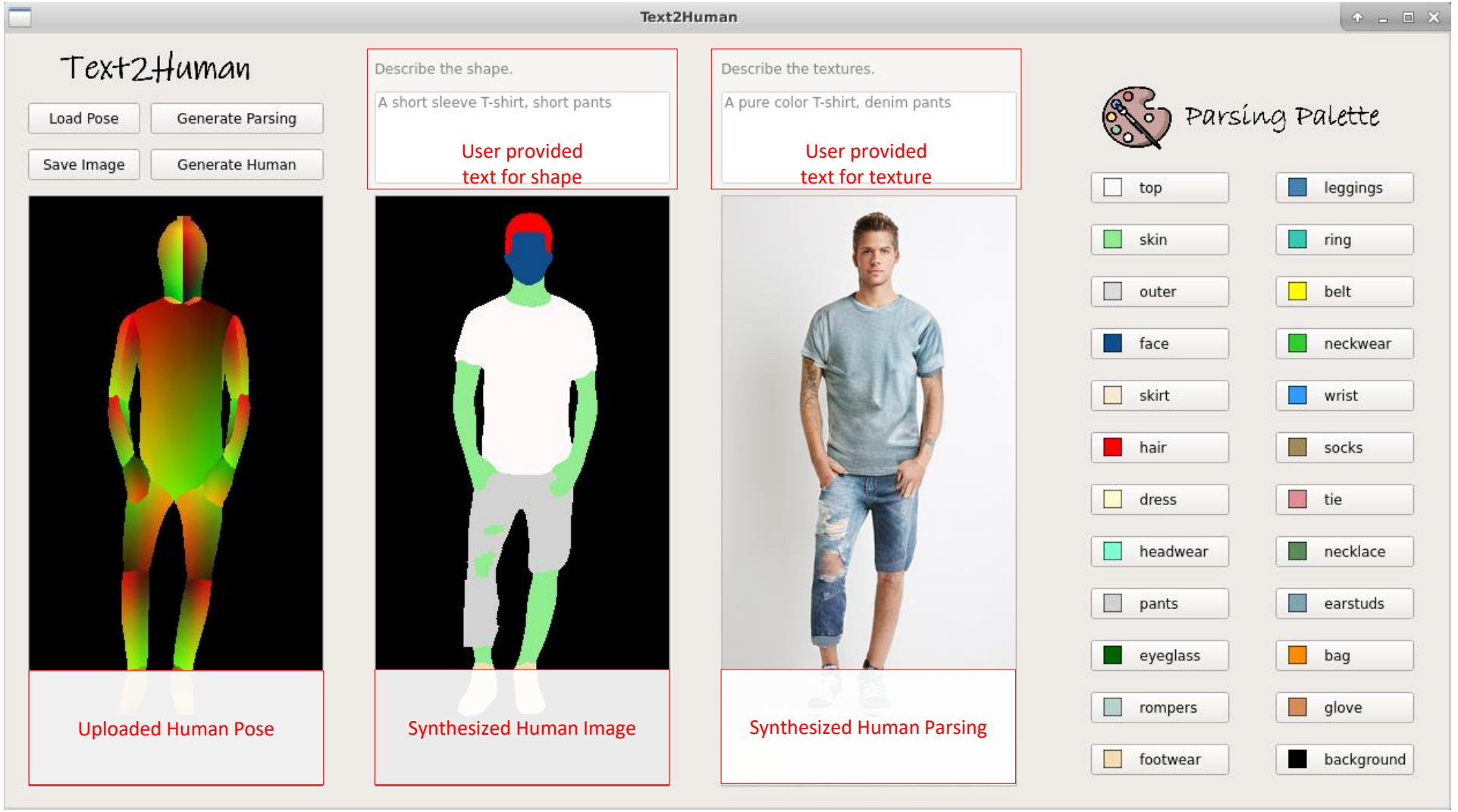
FRAMEWORK OF TEXT2HUMAN



FRAMEWORK OF TEXT2HUMAN



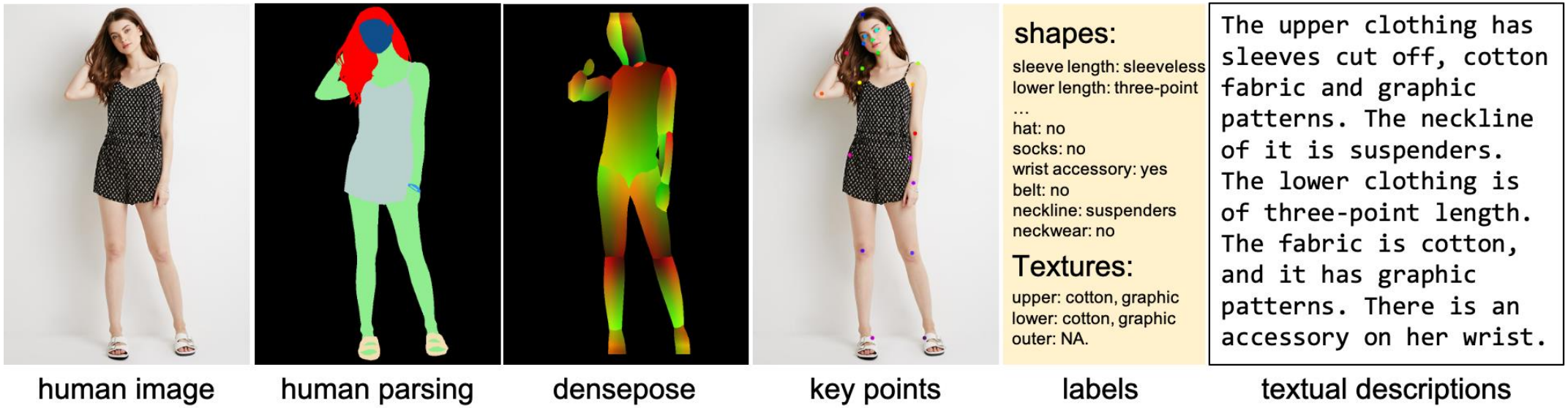
INTERACTIVE USER INTERFACE



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



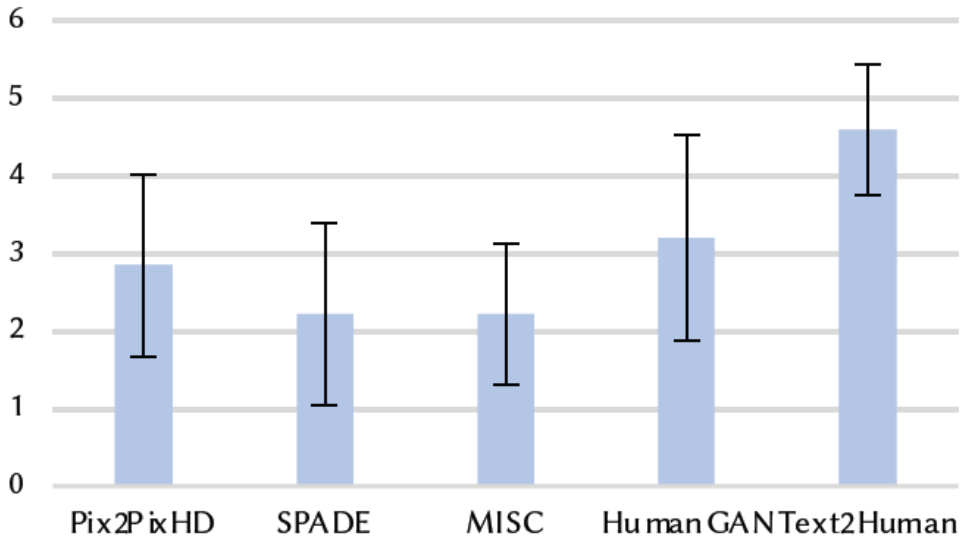
EXPERIMENT

pure color upper clothes with a denim outer, seven-point and pure color pants

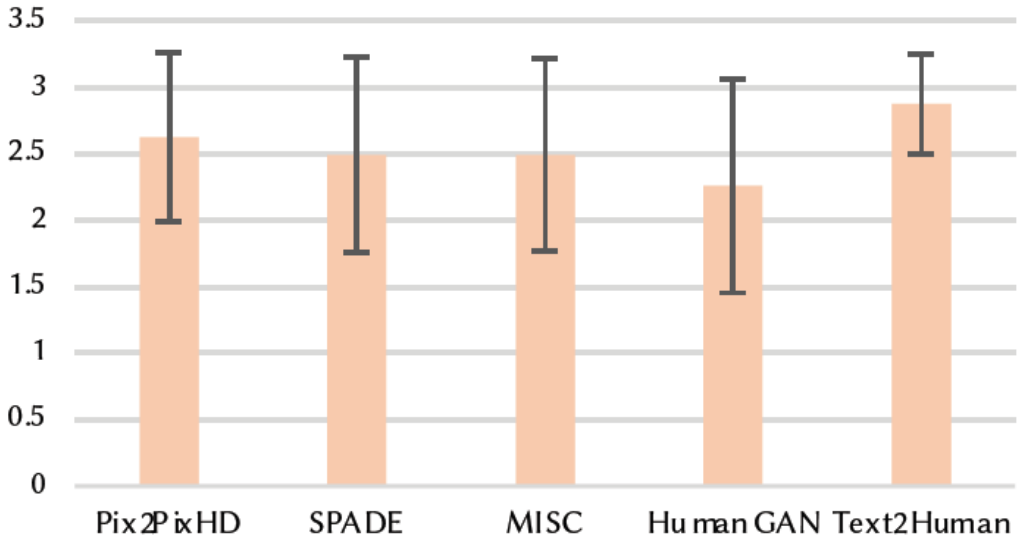
floral upper clothes with a pure color outer, three-point jeans

Parsing | Pix2PixHD | SPADE | MISC | HumanGAN | Text2Human

EXPERIMENT



(a) photorealism



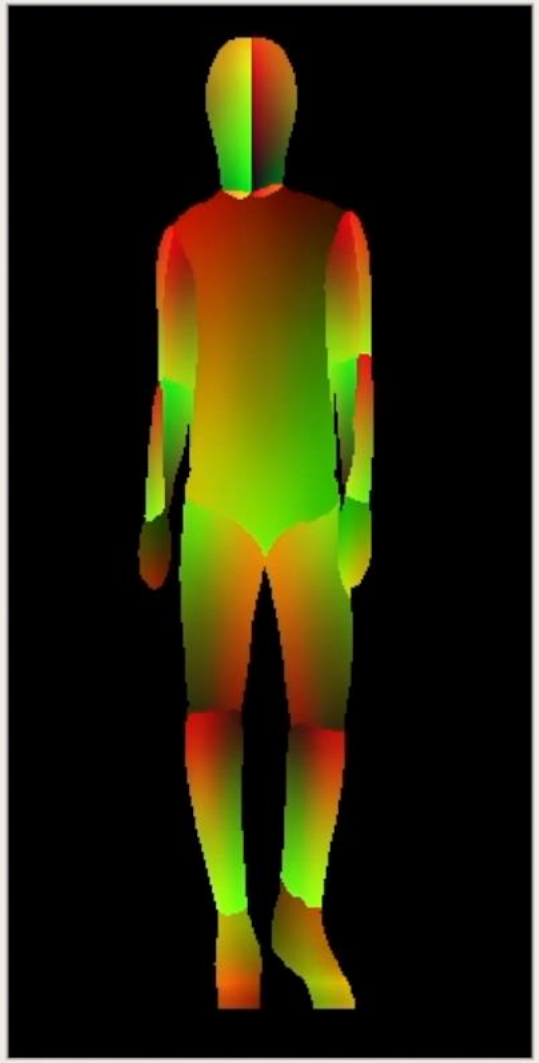
(b) texture consistency score

Text2Human

Text2Human

Load Pose | Generate Parsing

Save Image | Generate Human



Describe the shape.

Describe the textures.

Modify the desired texture by texts.

Parsing Palette

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.

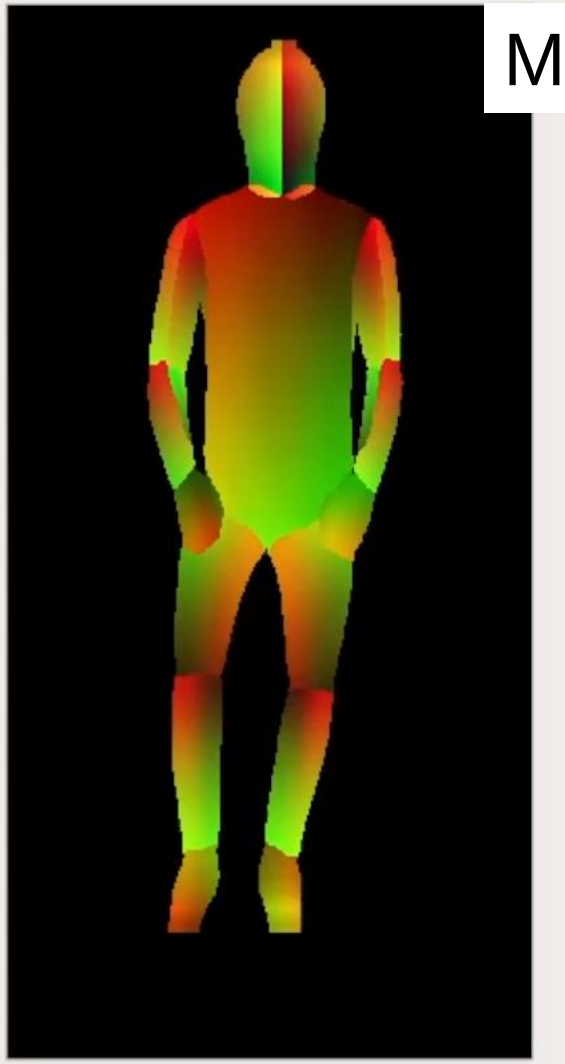
Load

Save Image

Generate Human

Manually add more details on the generate human parsing.

Modify the desired clothes shape by texts.



Add clothes attributes

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







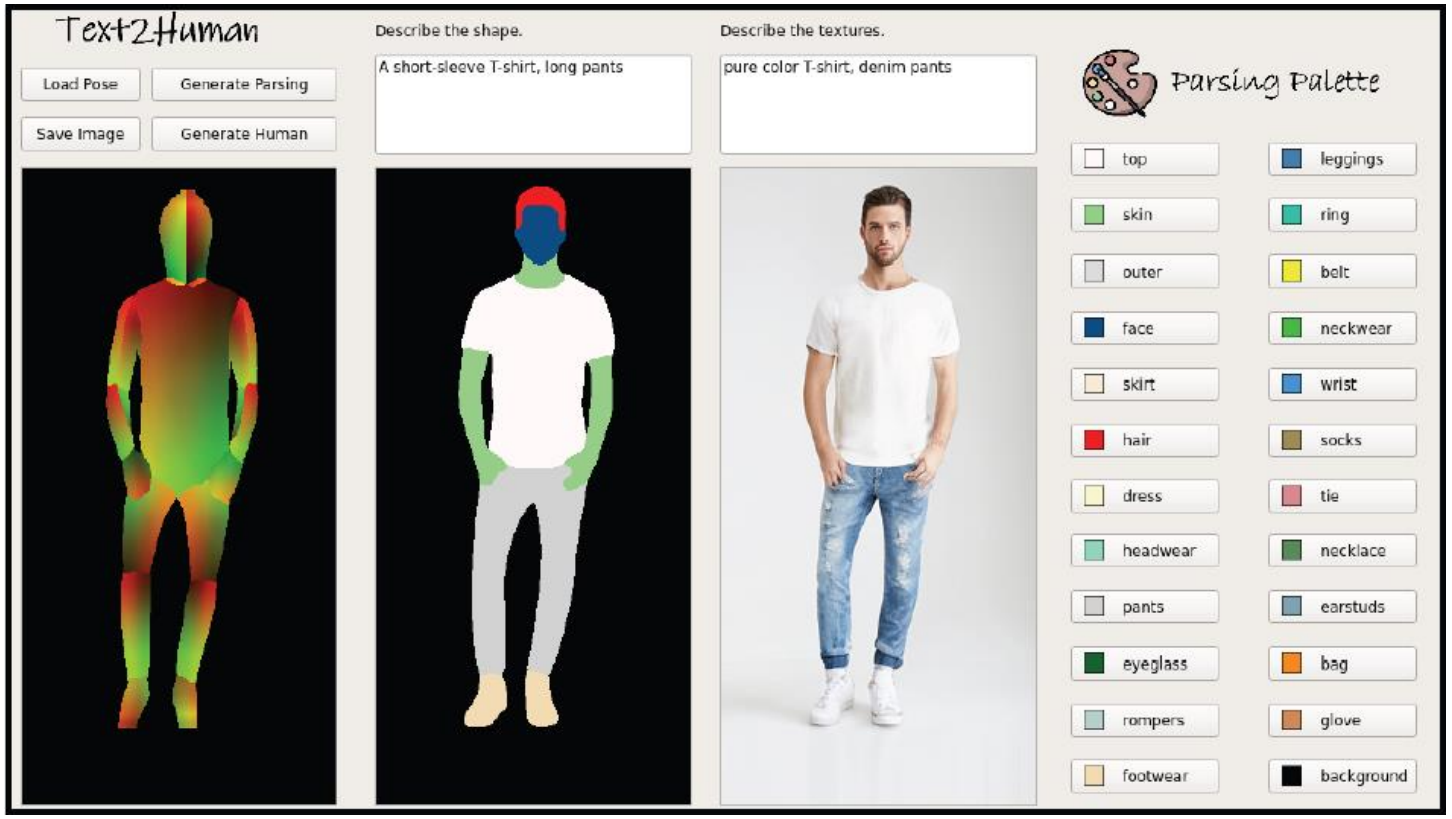






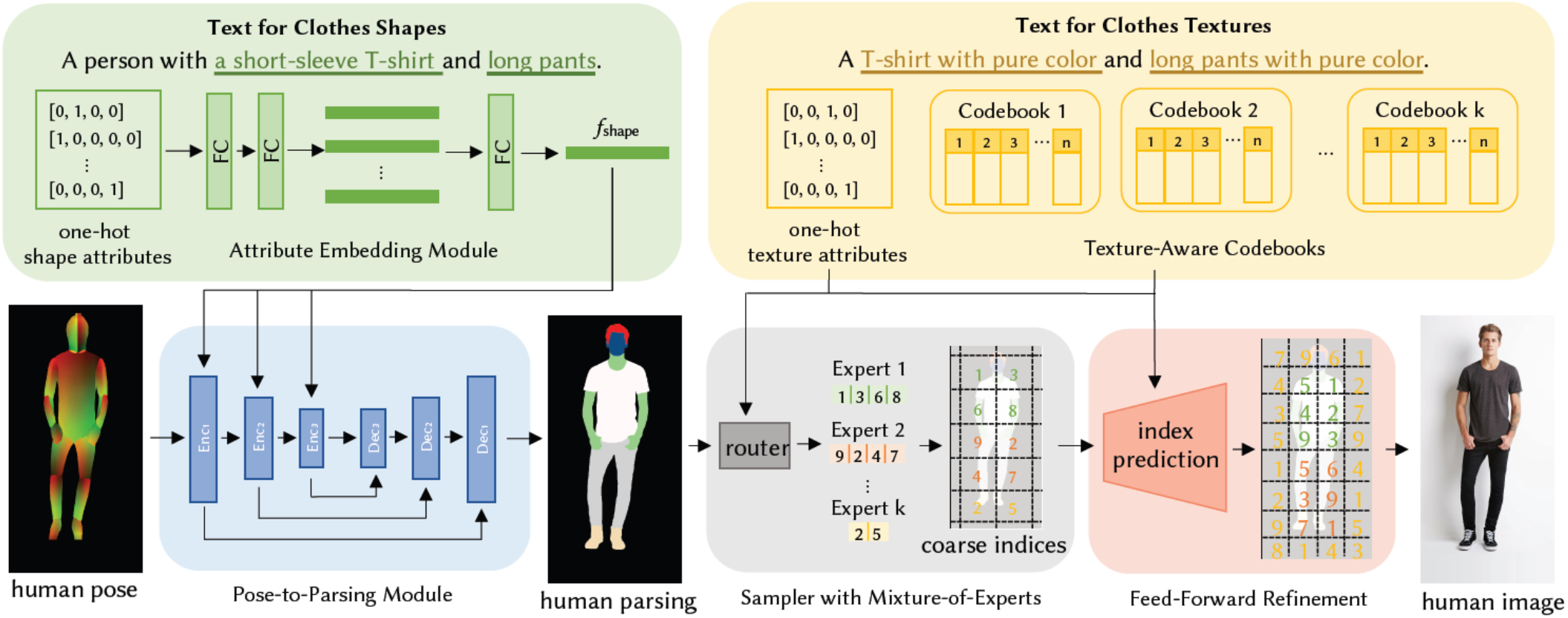
Task

Controllable Human Image Generation



Method

Text2Human



Dataset

DeepFashion-Multimodal

CODE AND MODELS

2D Generation

3D Generation

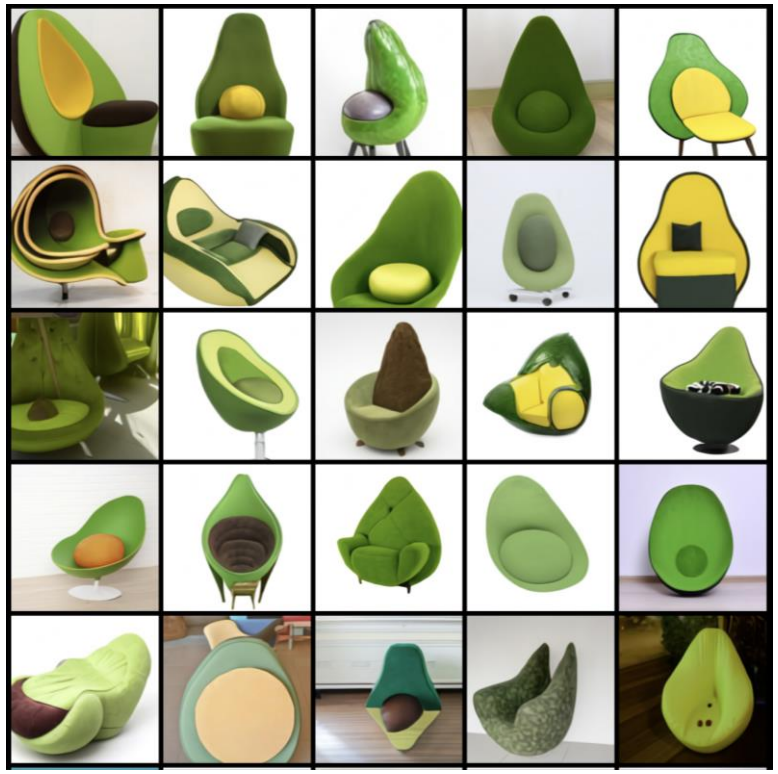
Data

Software

Motion Generation

Scene Generation

TEXT-DRIVEN IMAGE GENERATION



DALL·E [1]

DALL·E 2 [2]

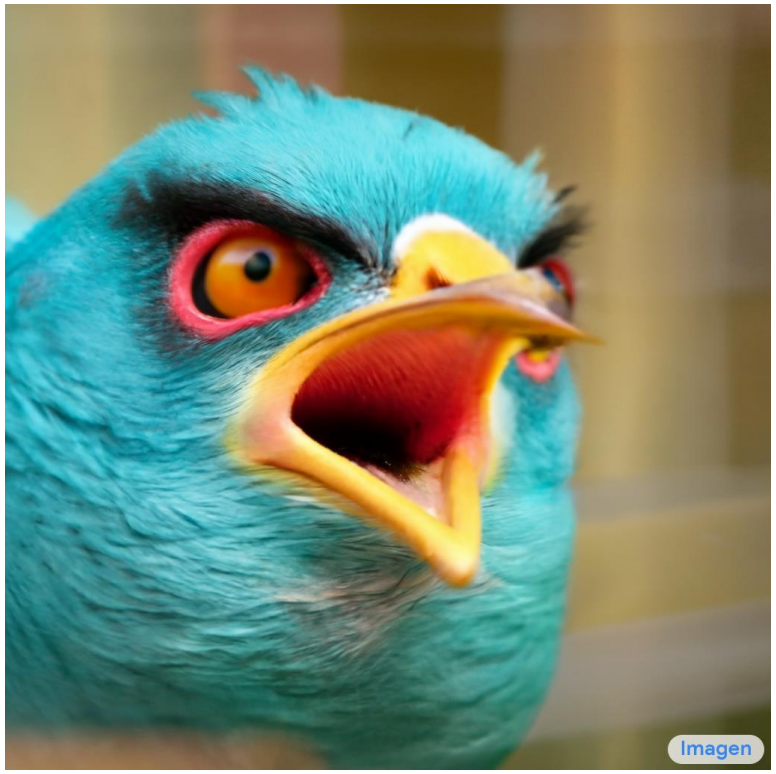


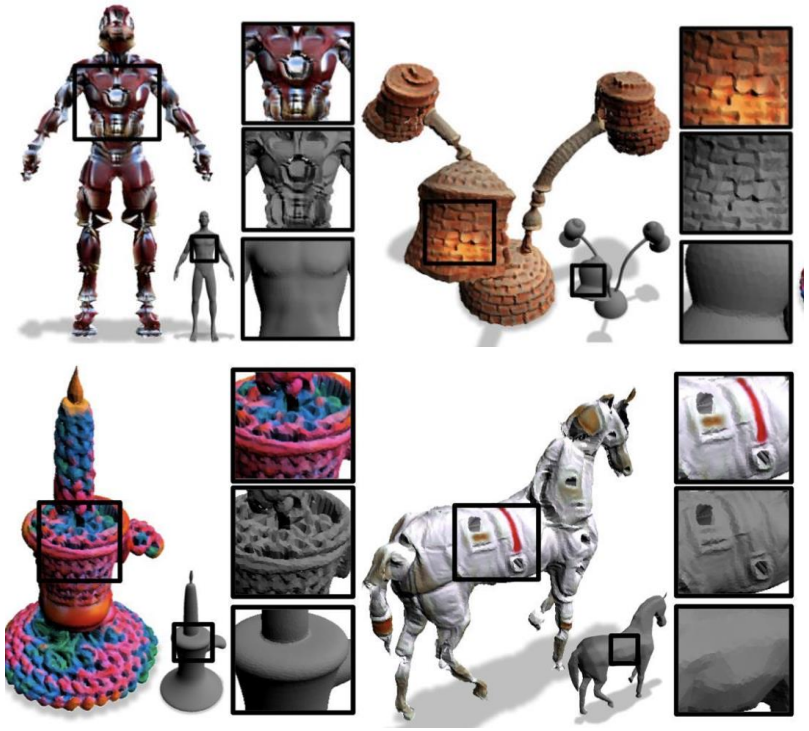
Imagen [3]

[1] https://openai.com/blog/dall-e/ [2] https://openai.com/dall-e-2/
[3] https://imagen.research.google

TEXT-DRIVEN 3D GENERATION

CLIP + DIFFERENTIABLE RENDERING

Dream Field [1]



Text2Mesh [2]

[1] https://ajayj.com/dreamfields [2] https://threedle.github.io/text2mesh/

WHAT ABOUT TEXT-DRIVEN AVATAR GENERATION => NOW WE HAVE AVATARCLIP

I want to generate a tall and fat Iron Man that is running.

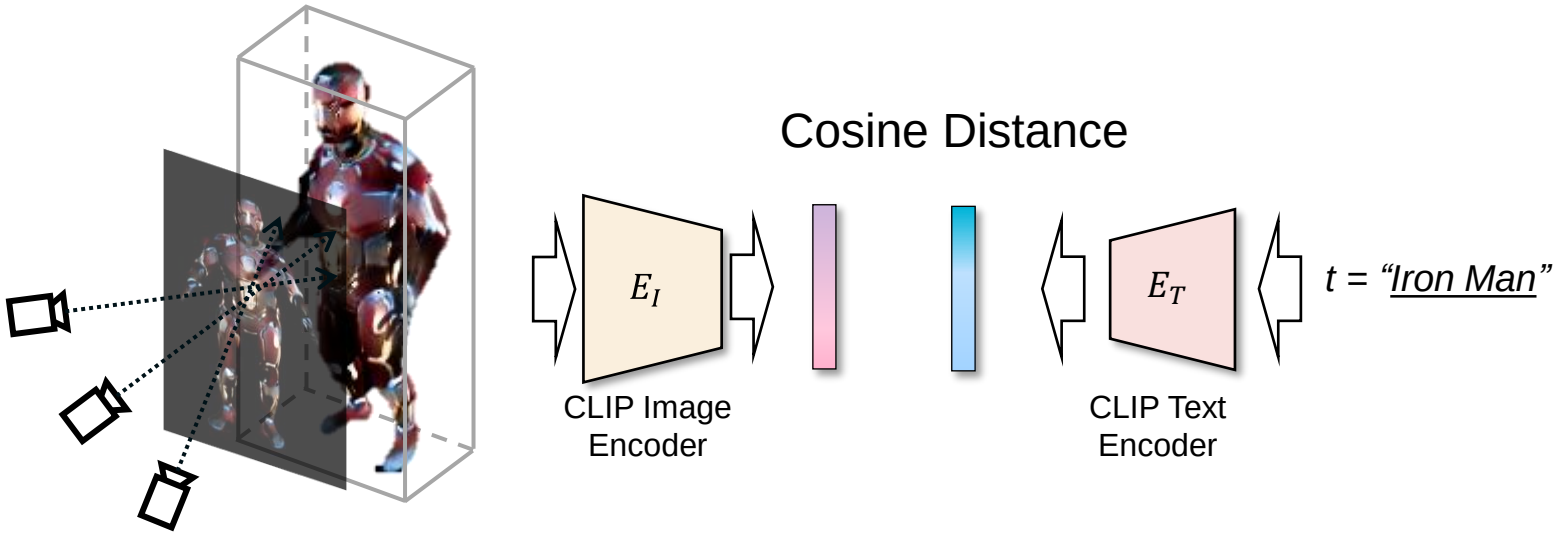
I would like to generate a skinny ninja that is raising arms.

I want to generate a tall and skinny female soldier that is arguing.

I want to generate an overweight sumo wrestler that is sitting.

TEXT-DRIVEN 3D GENERATION

CLIP + DIFFERENTIABLE RENDERING



a) Differentiable Rendering

b) Optimization guided by CLIP

AVATARCLIP: HOW IT WORKS

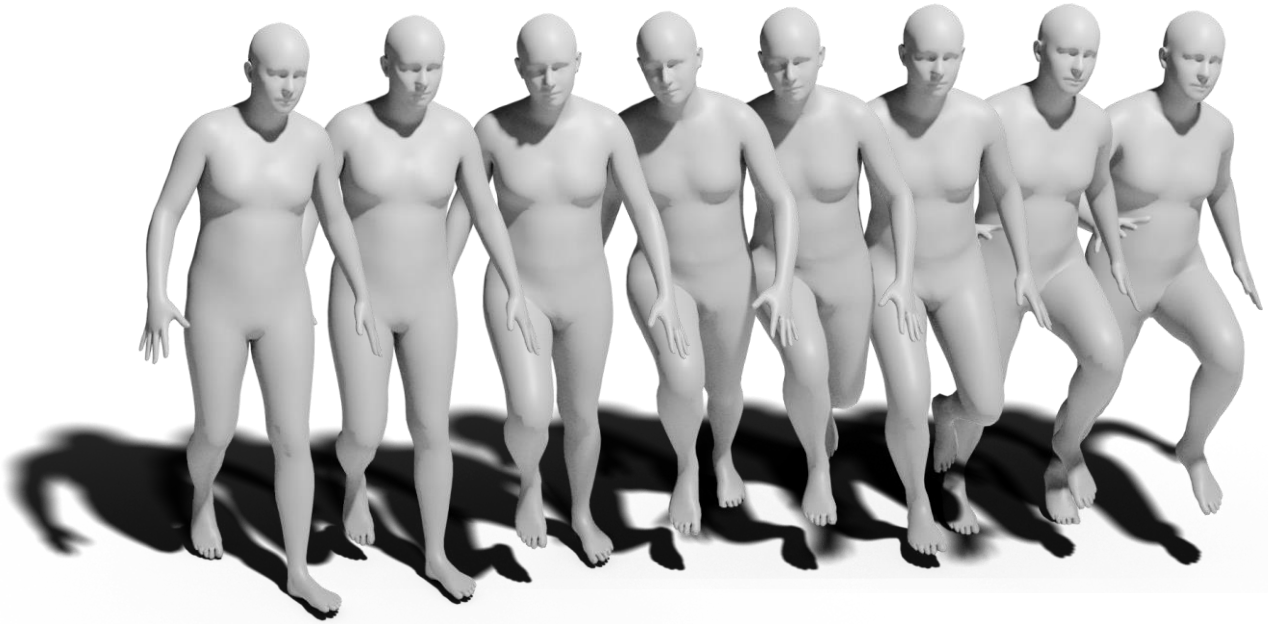
A) STATIC AVATAR GENERATION

Shape Description: *“a tall and fat man”*

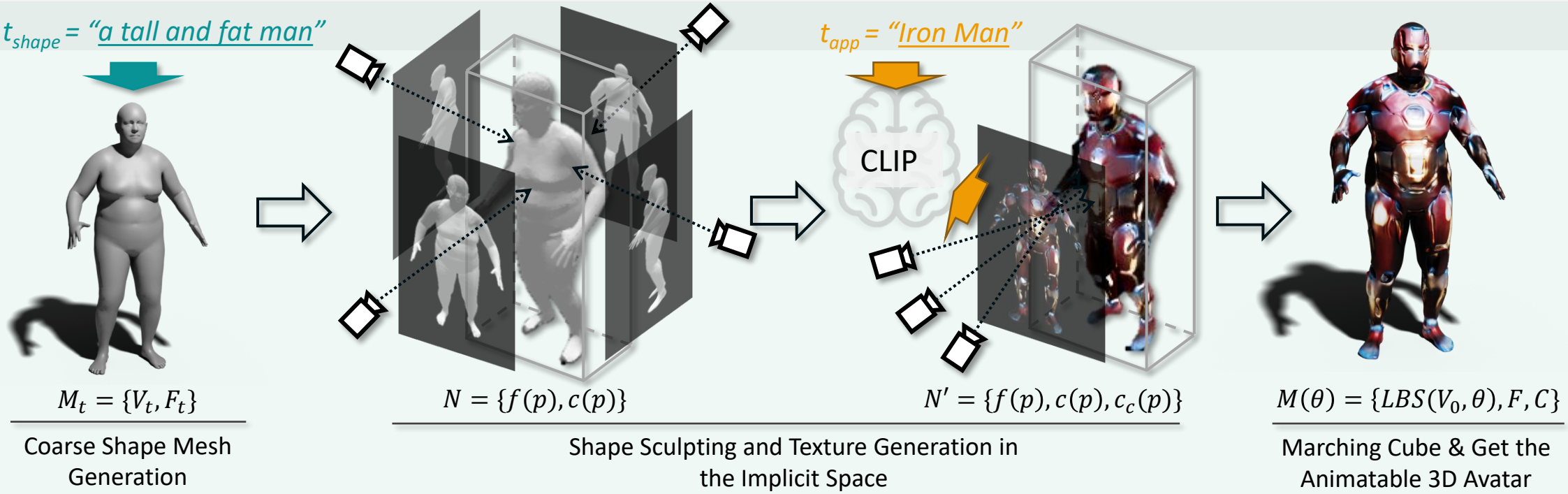
Appearance Description: *“Iron Man”*

B) MOTION GENERATION

Motion Description: *“running”*

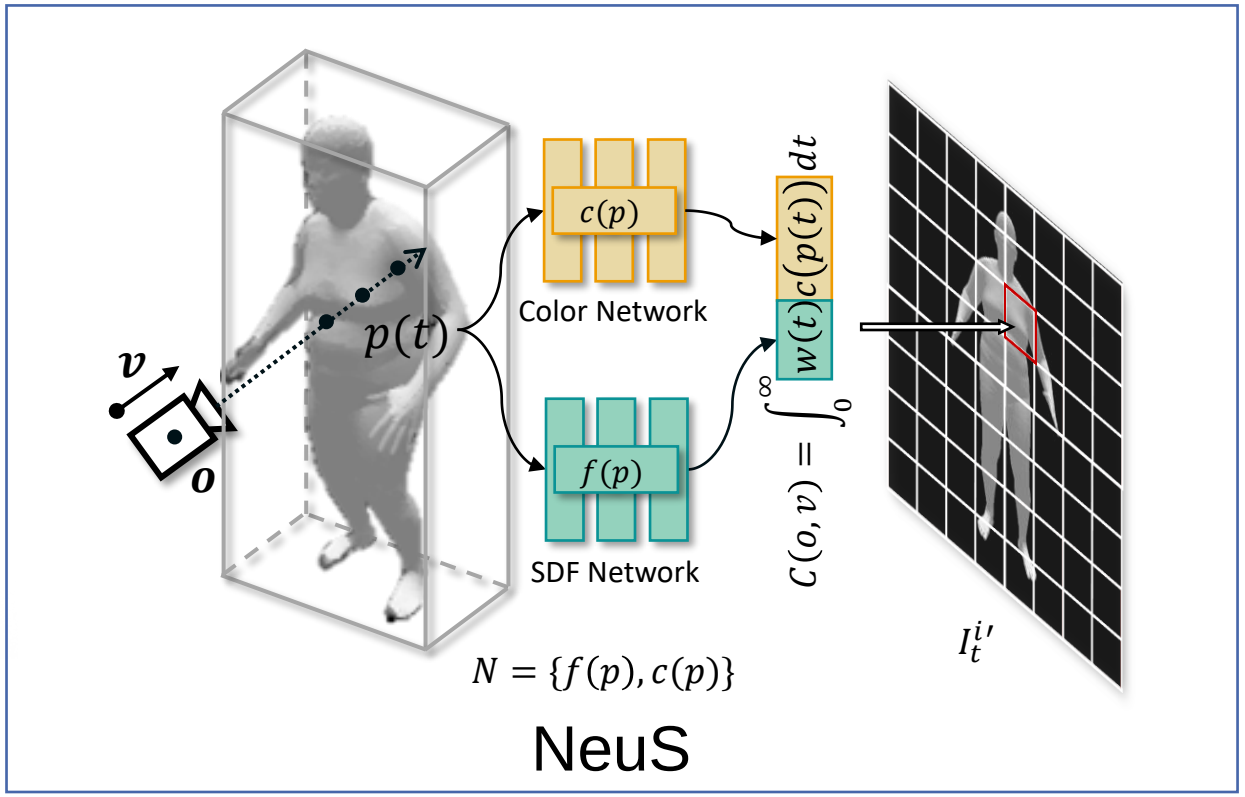


AVATARCLIP: DETAILED PIPELINE

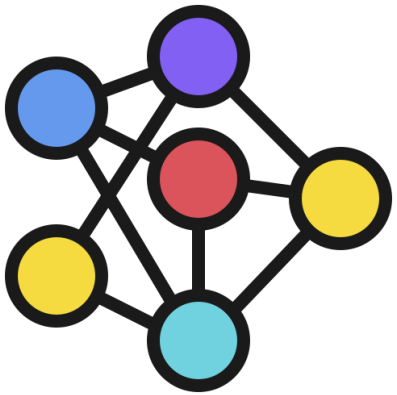


AVATARCLIP: TO THE IMPLICIT SPACE

Mesh

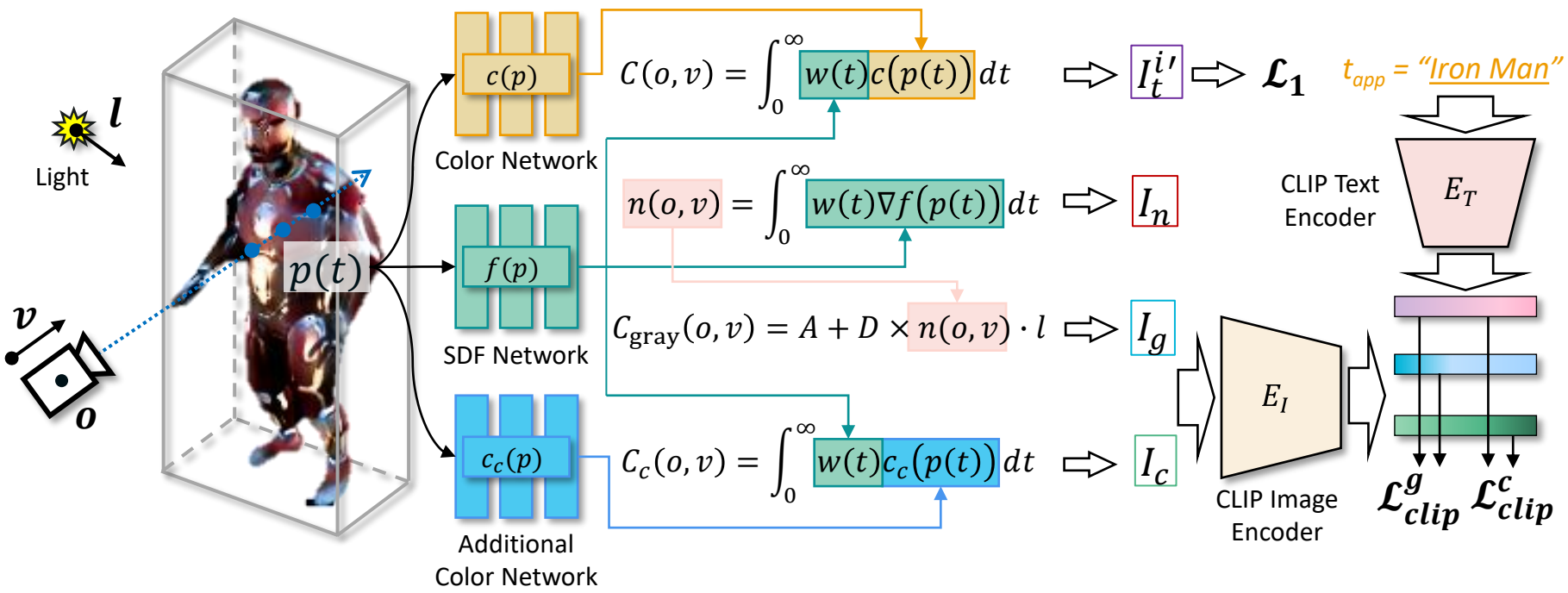


NeuS



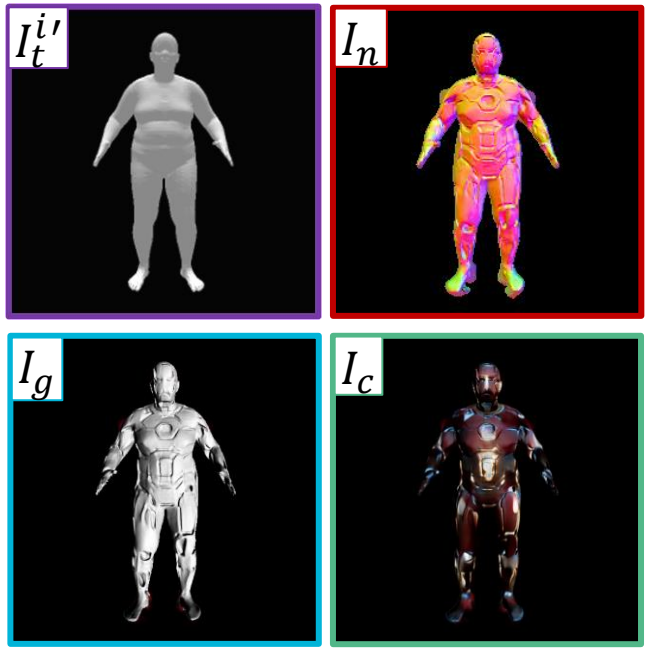
Implicit Function

AVATARCLIP: SHAPE SCULPTING AND TEXTURE GENERATION



a) Rendering the Implicit 3D Avatar $N' = \{f(p), c(p), c_c(p)\}$

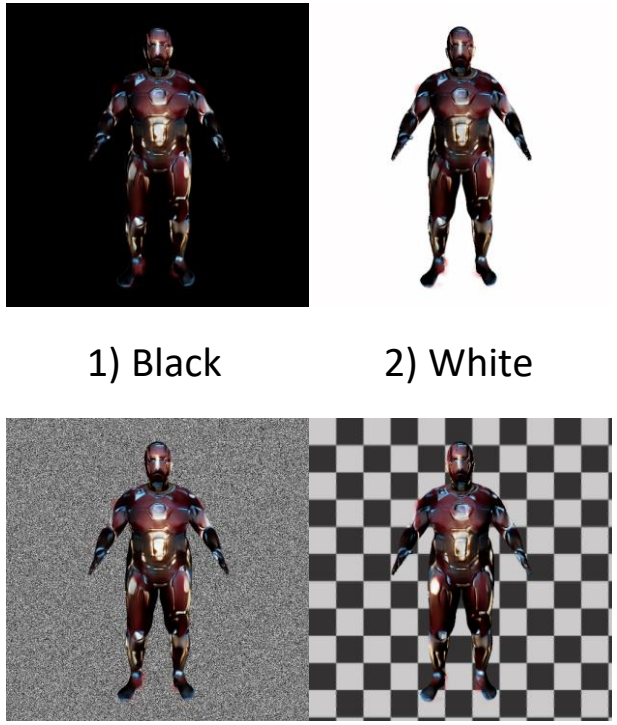
b) Optimization



Examples of Intermediate Results

AVATARCLIP: OPTIMIZATION PROCESS

A) RANDOM BACKGROUND SEGMENTATION



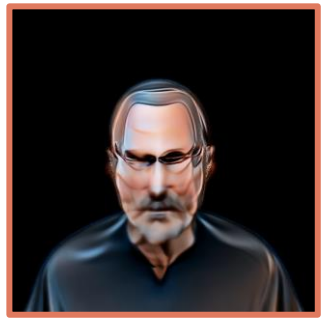
1) Black

2) White

3) Gaussian Noise

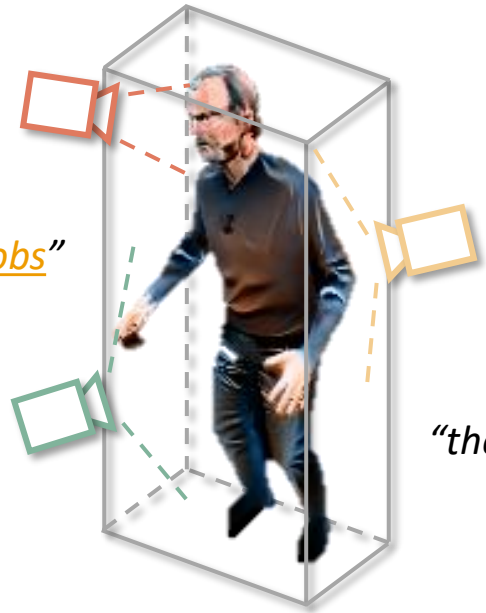
4) Chess Board

B) SEMANTIC-AWARE PROMPT AUGMENTATION



"the face of Steve Jobs"

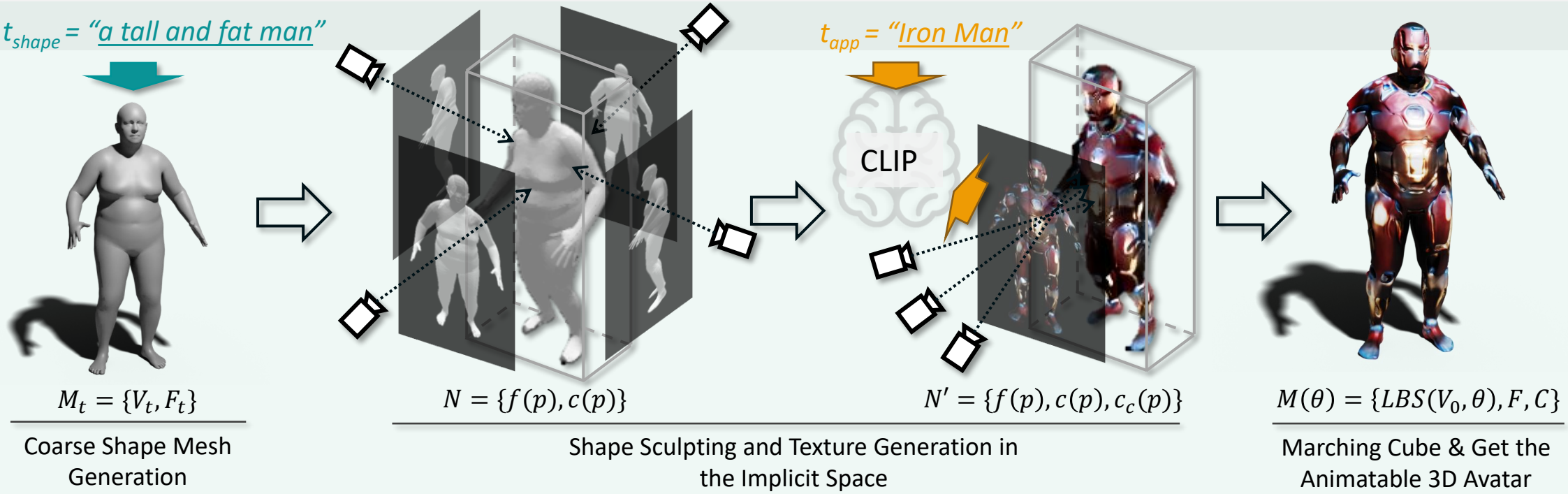
"Steve Jobs"



"the back of Steve Jobs"

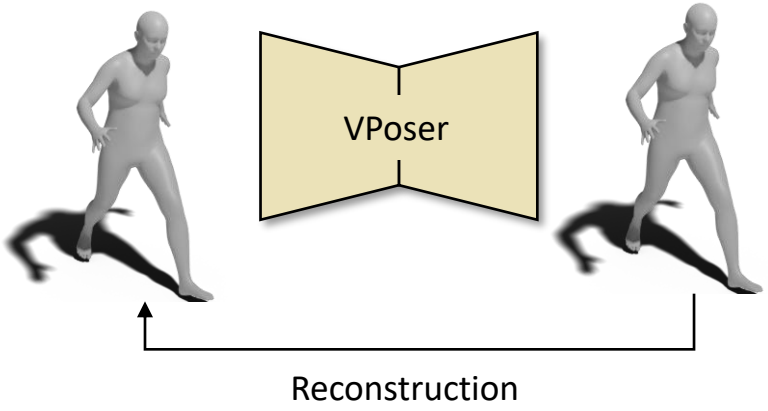
Implicit 3D Avatar $N' = \{f(p), c(p), c_c(p)\}$

AVATARCLIP: DETAILED PIPELINE

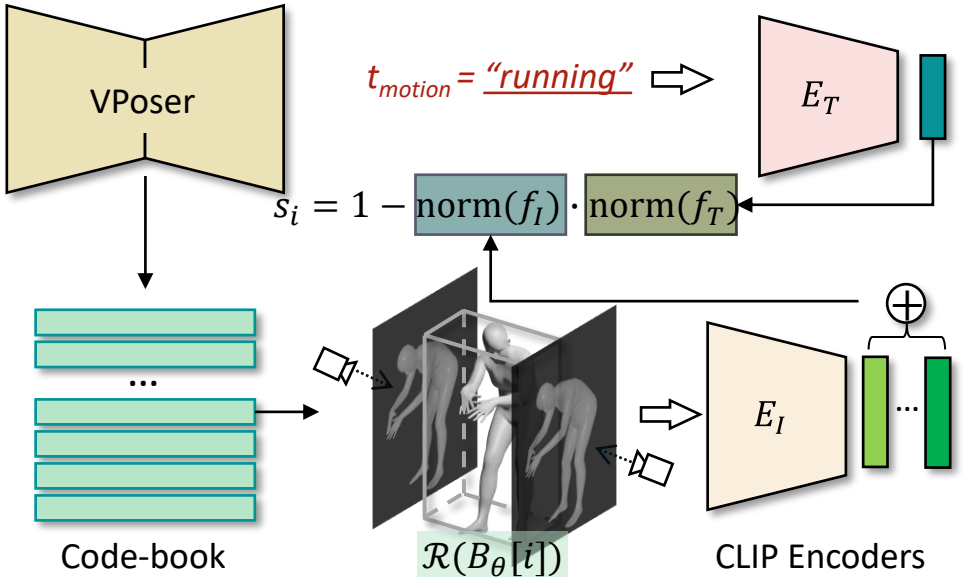


AVATARCLIP: CANDIDATE POSES GENERATION

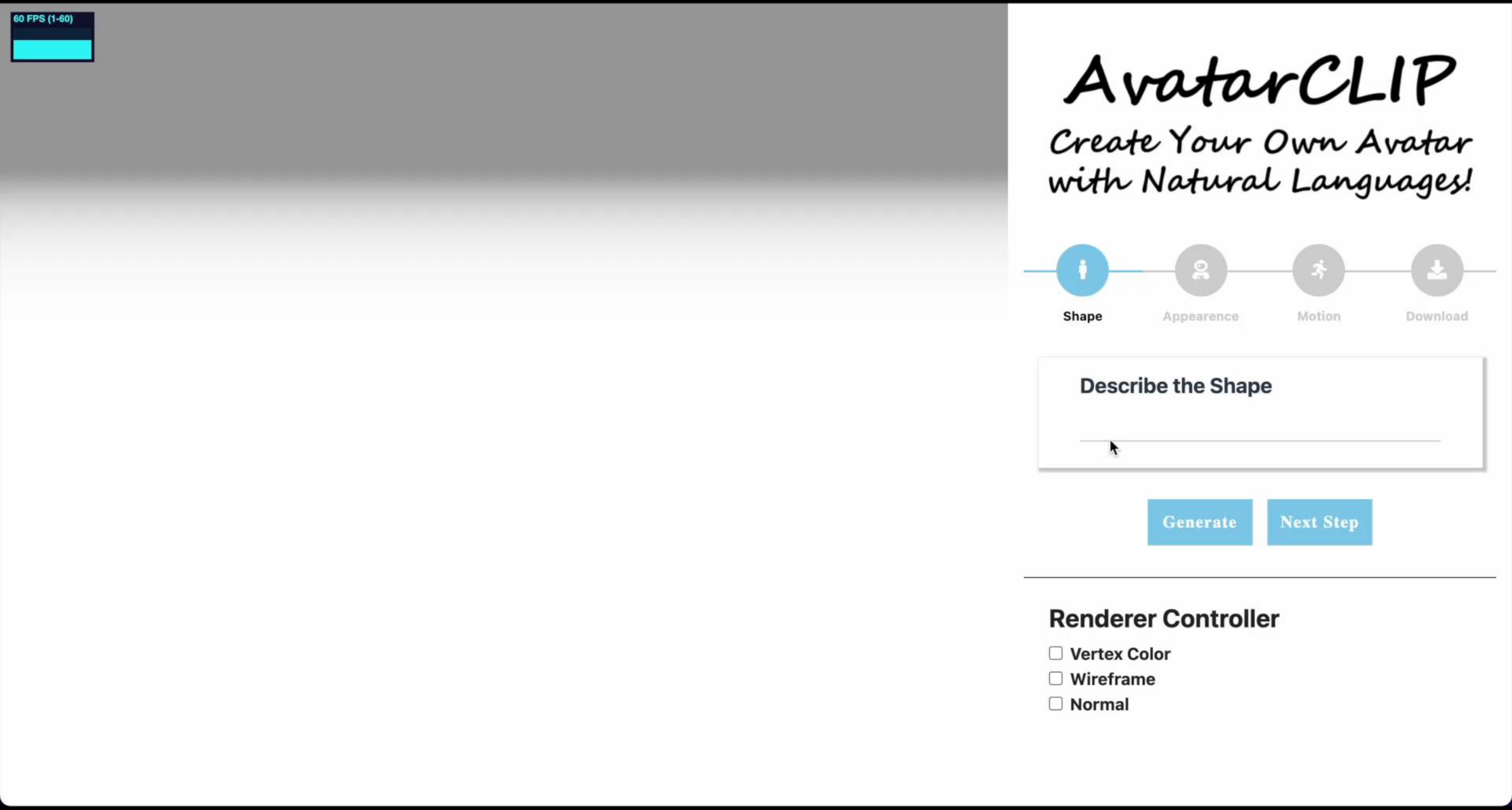
A) POSE VAE (VPOSER)



B) CLIP-GUIDED CANDIDATE POSES QUERY

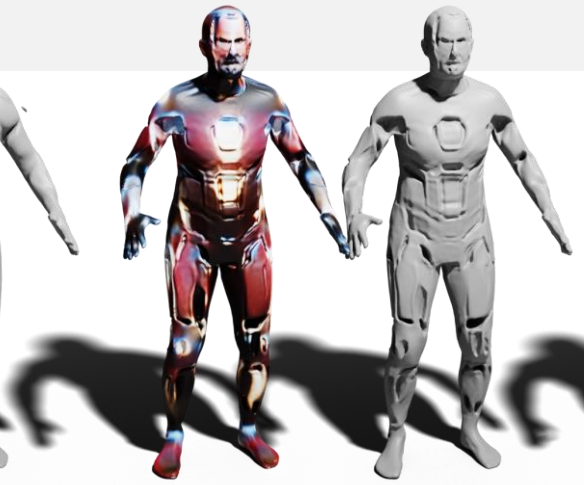


OVERALL RESULTS


60 FPS (1-60)
AvatarCLIP
Create Your Own Avatar with Natural Languages!
Shape
Appearence
Motion
Download
Describe the Shape
Generate
Next Step
Renderer Controller
Vertex Color
Wireframe
Normal

CONTROLLING & CONCEPT MIXING ABILITIES

1. Superman
2. the face of Bill Gates



1. Iron Man
2. the face of Steve Jobs

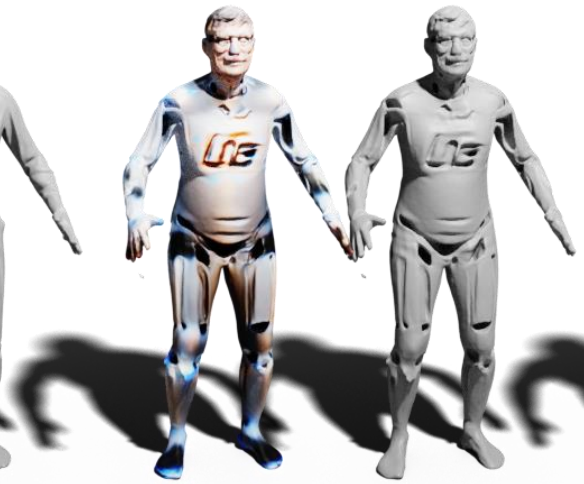
Steve Jobs in White Shirt

Man in Jeans

Man in White Shirt

Alien Bill Gates

Bill Gates Wearing Batman Suit



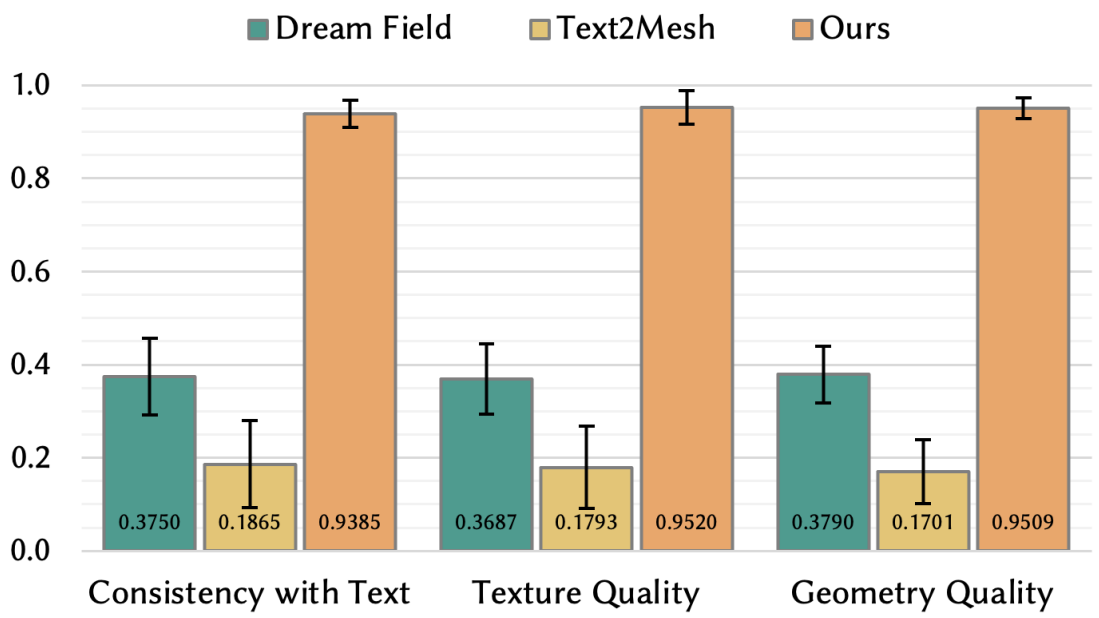
Robot Bill Gates

Zombie Steve Jobs

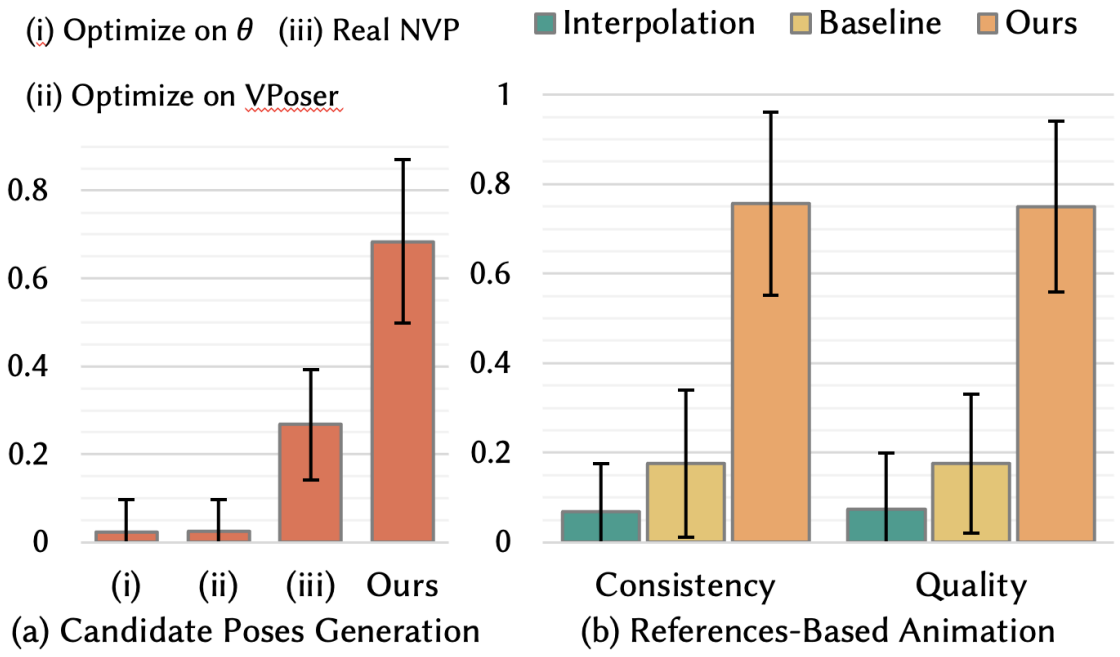
Zombie Iron Man

QUANTITATIVE RESULTS: USER STUDY

A) STATIC AVATAR GENERATION



B) MOTION GENERATION



CODE AND MODELS

GitHub

Project Page

2D Generation

3D Generation

Data

Software

Motion Generation

Scene Generation

3D Animation

Video Games

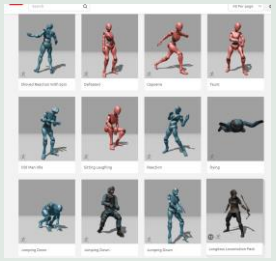
Films

VTuber

Motion Collection

Manual Editing

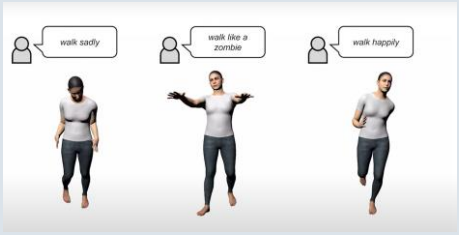
Motion Capture



Gallery

1. **Expensive**
2. **Time-consuming**
3. **Not User-friendly**

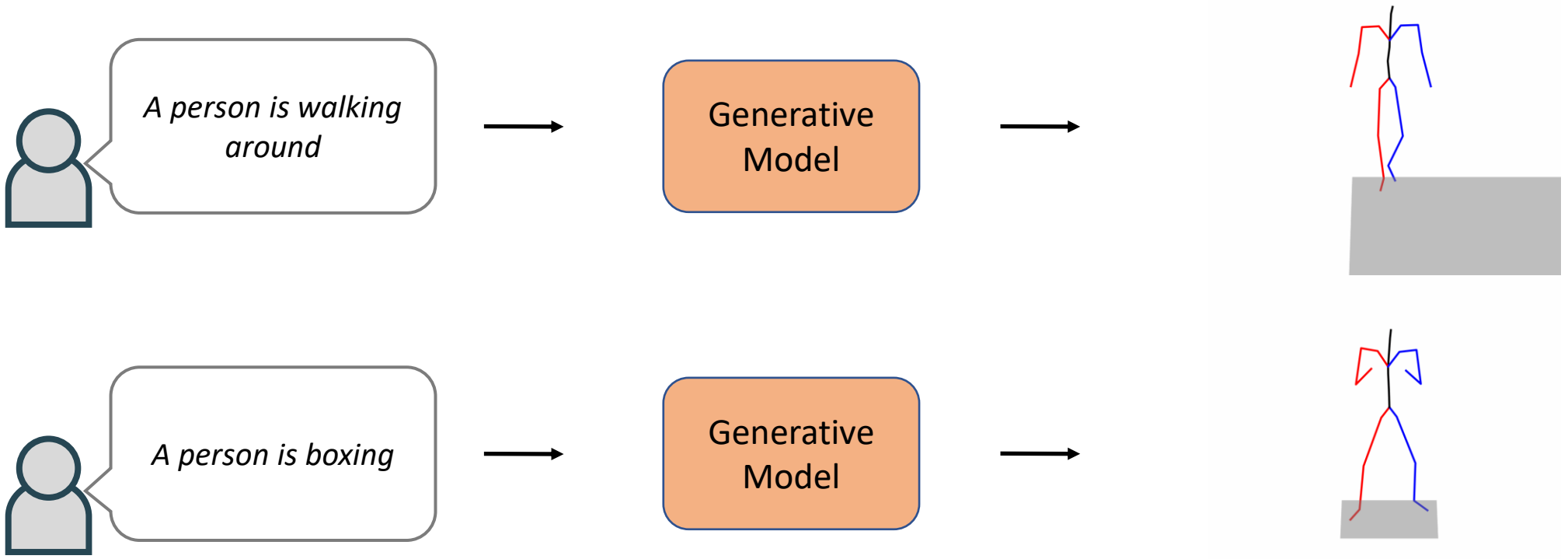
Human Mesh Recovery



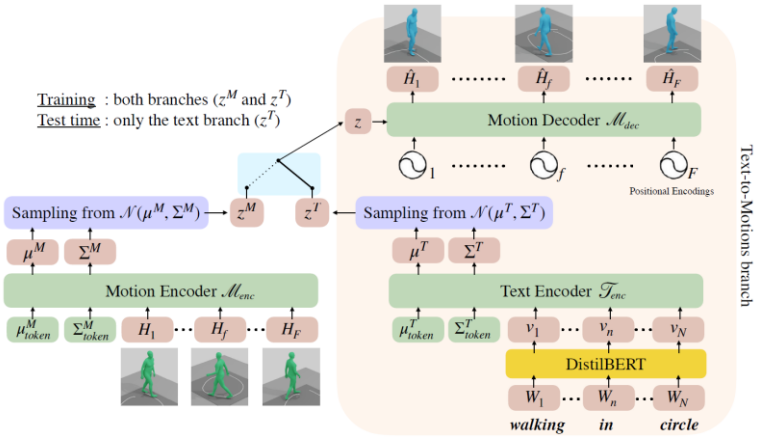
Conditional Motion Generation

1. **Cheap**
2. **Efficient**
3. **User-friendly**

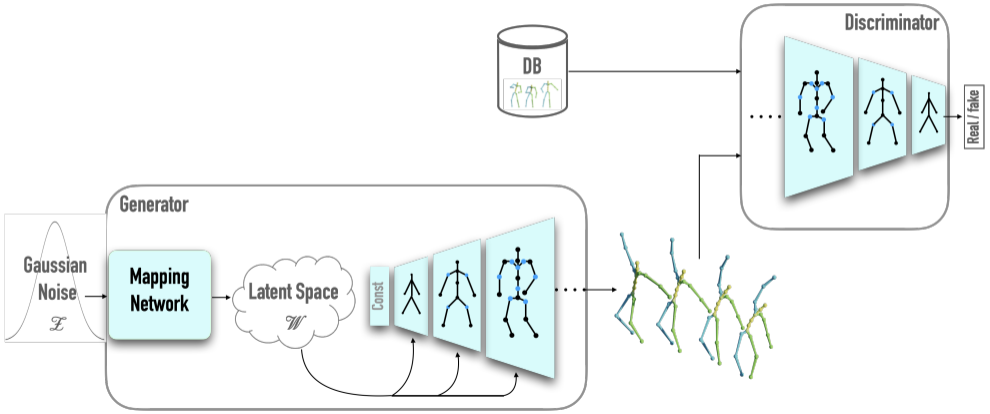
Text-driven Motion Generation



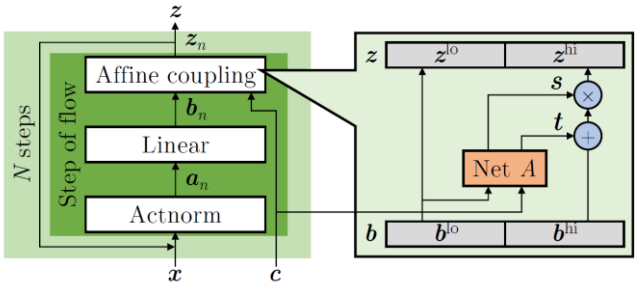
Classical Motion Generative Model



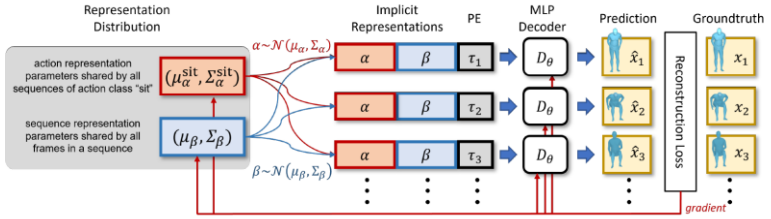
TEMOS[1] (VAE)



MoDi[2] (GAN)



MoGlow[3] (Normalization Flow)



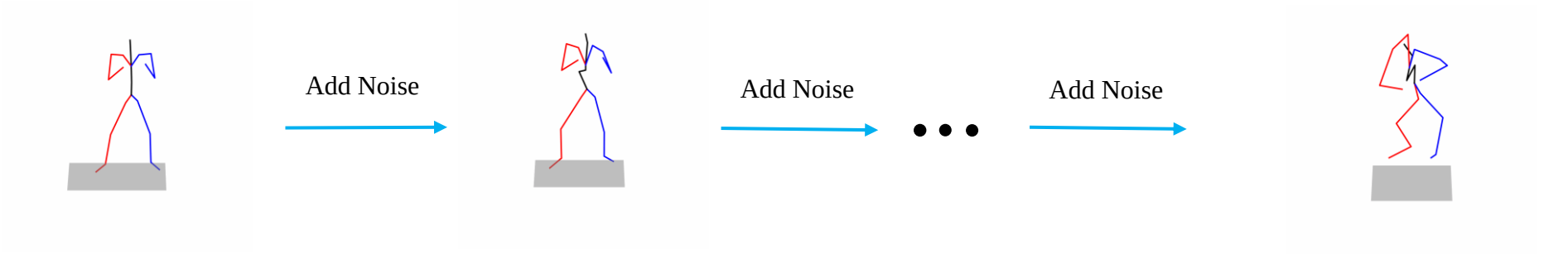
INR[4] (Implicit Function)

Issues: 1) Hard to model complicated motion sequence 2) Lack of diversity

[1] Petrovich M, et al. Temos: Generating diverse human motions from textual descriptions. ECCV 2022
[2] Sigal R, et al. MoDi: Unconditional Motion Synthesis from Diverse Data. ArXiv 2022
[3] Henter GE, et al. Moglow: Probabilistic and controllable motion synthesis using normalising flows. TOG 2020
[4] Cervantes P, et al. Implicit neural representations for variable length human motion generation. ECCV 2022

Motion Generation with Diffusion Model

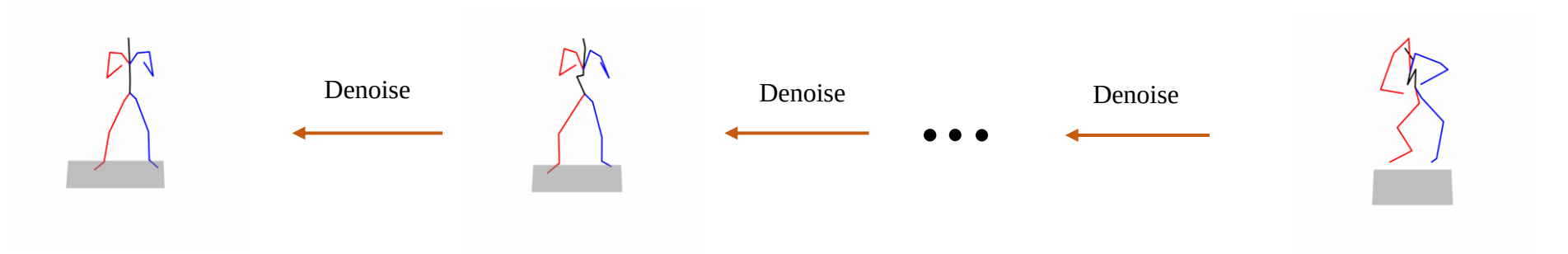
Diffusion Process



$\mathbf{x_0} \sim q(\mathbf{x_0})$

$p(\mathbf{x_T}) = \mathcal{N}(\mathbf{x_T}; \mathbf{0}, \mathbf{I})$

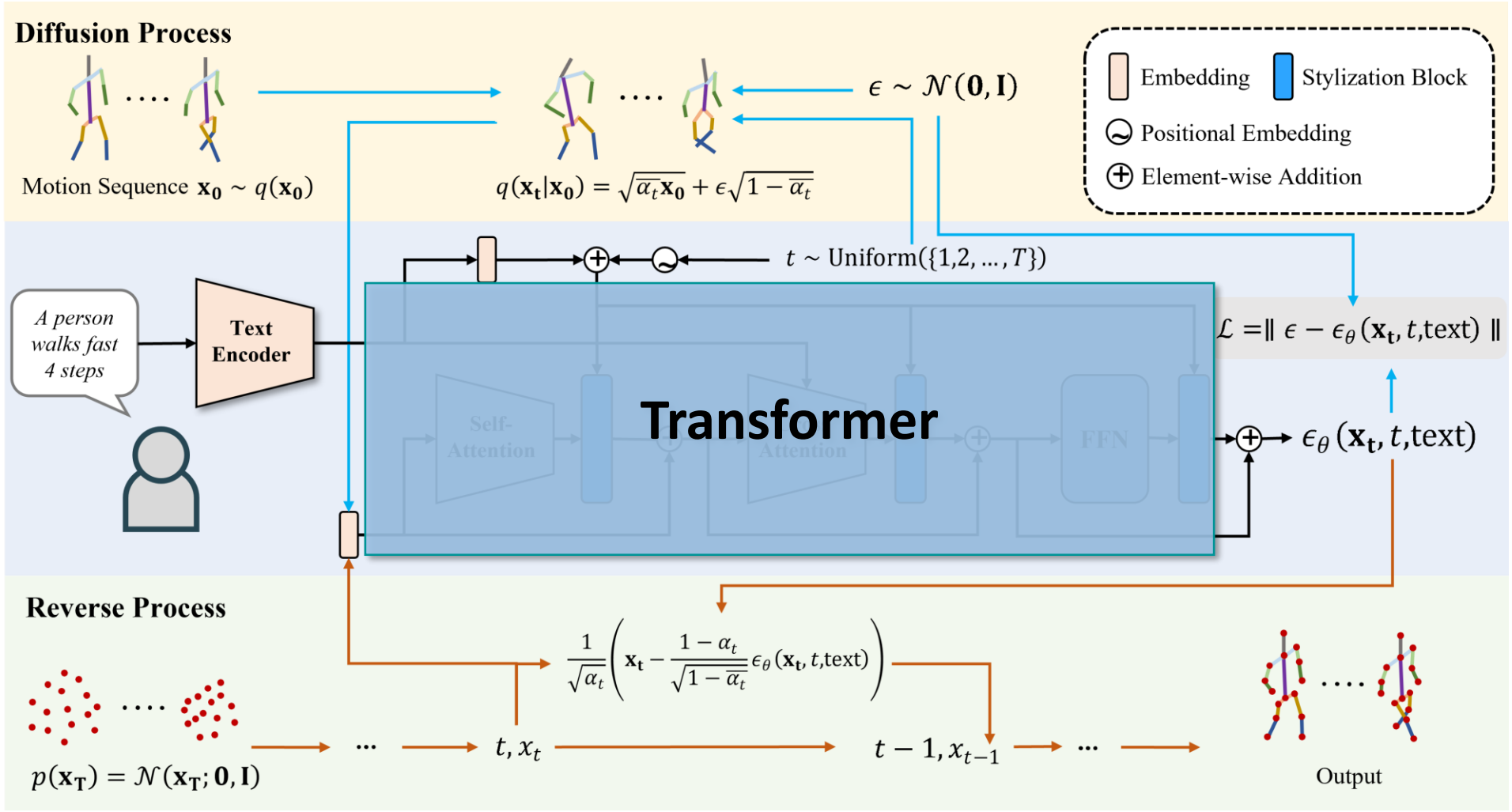
Reverse Process



$\mathbf{x_0} \sim q(\mathbf{x_0})$

$p(\mathbf{x_T}) = \mathcal{N}(\mathbf{x_T}; \mathbf{0}, \mathbf{I})$

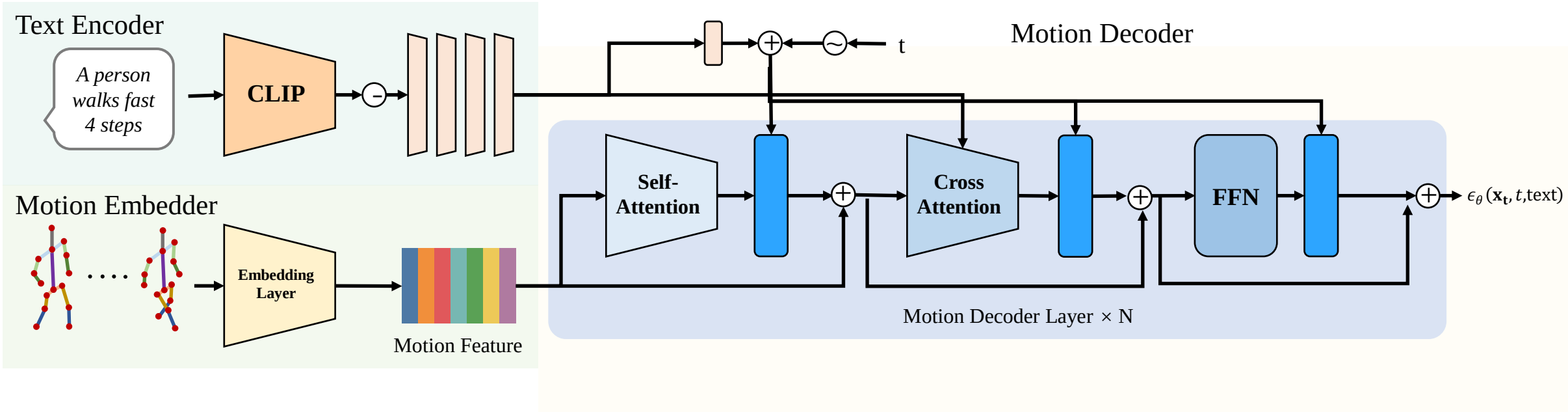
Framework



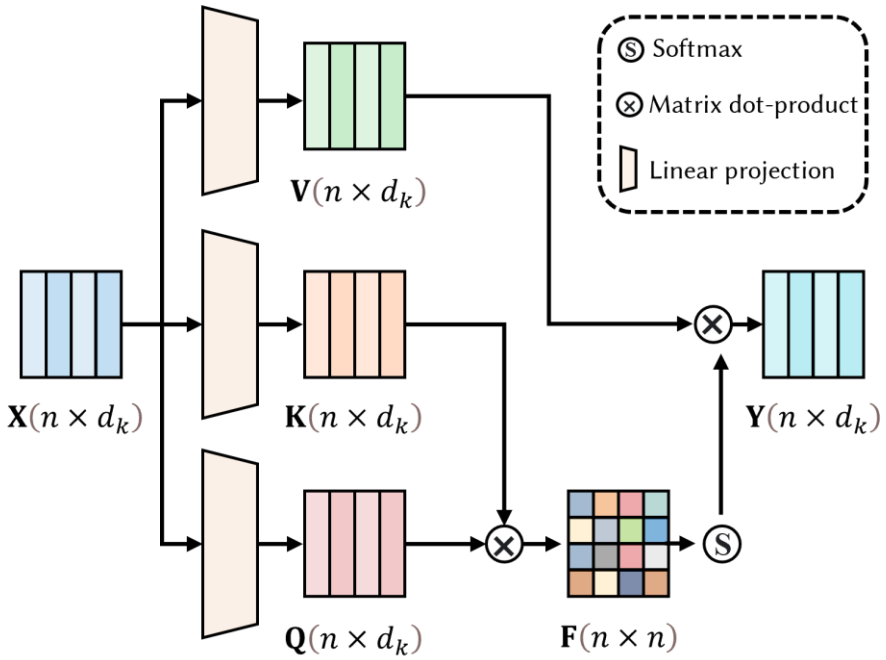
Challenge:

1. Variable length
2. Fusing timestep
3. Improve efficiency

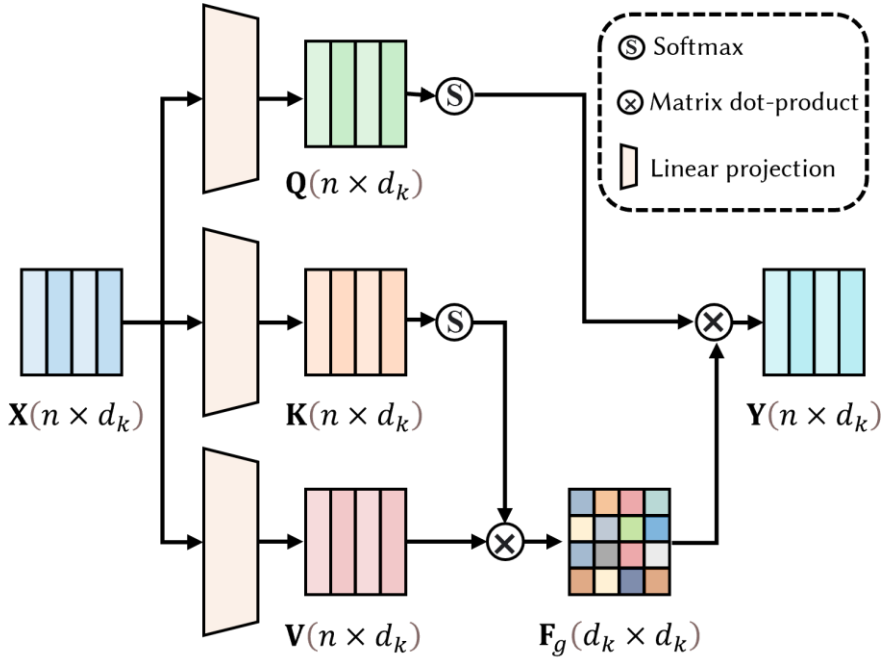
Cross-Modality Linear Transformer



Linear Self-Attention



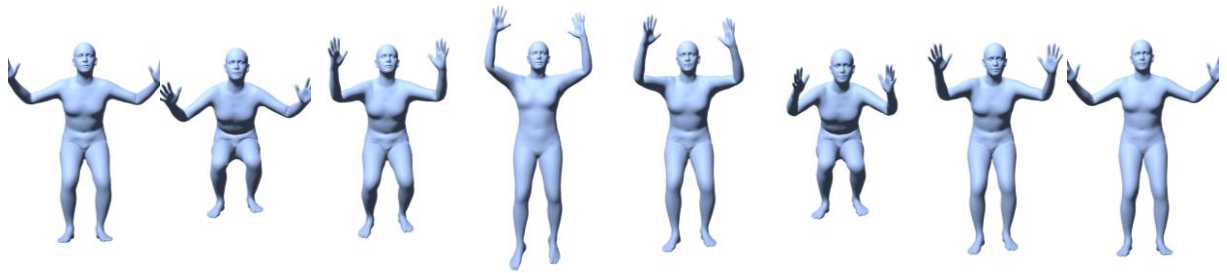
Classical Self-Attention



Linear Self-Attention

Visualization

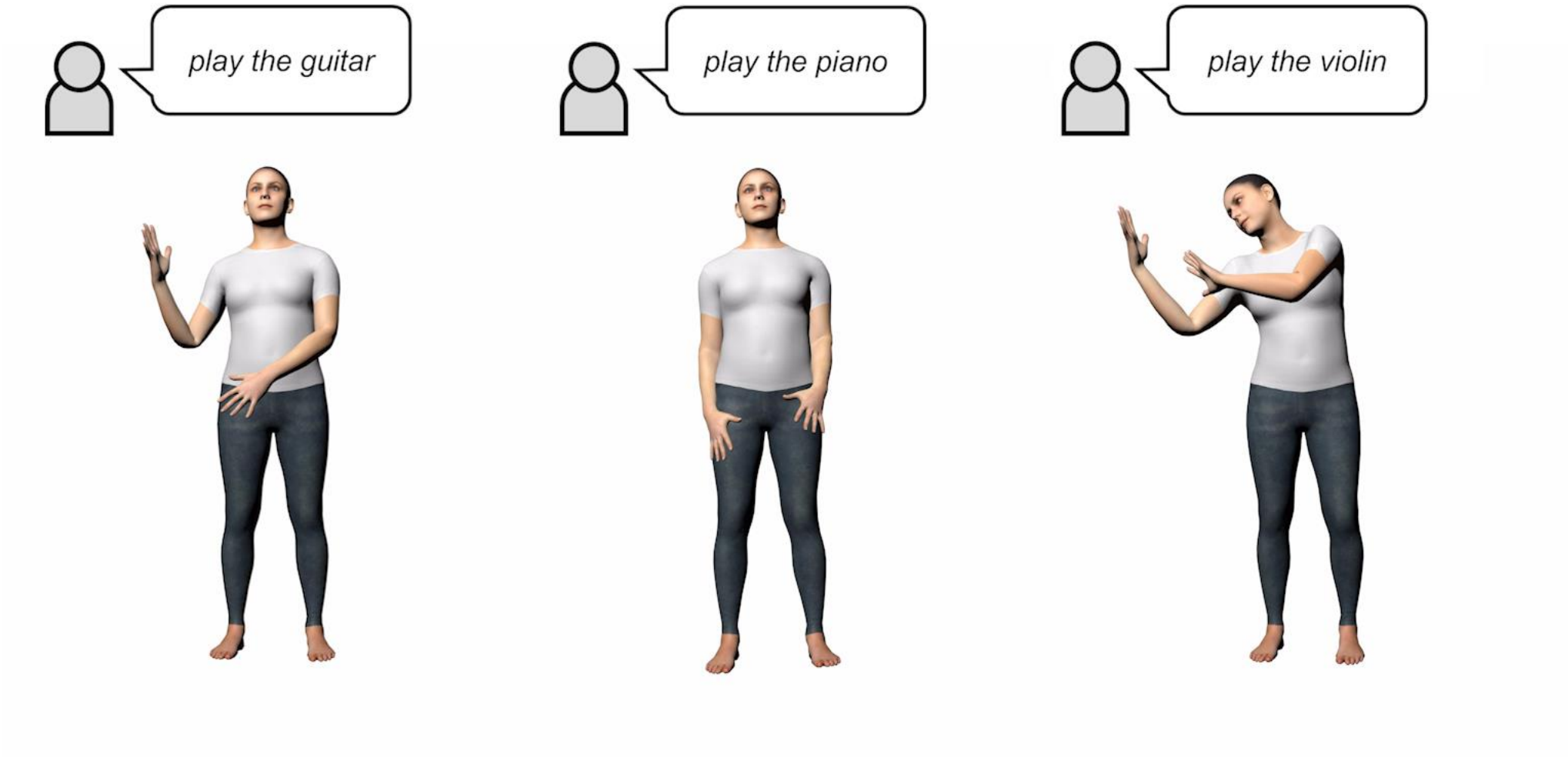
a) *Kicking* and *punching*



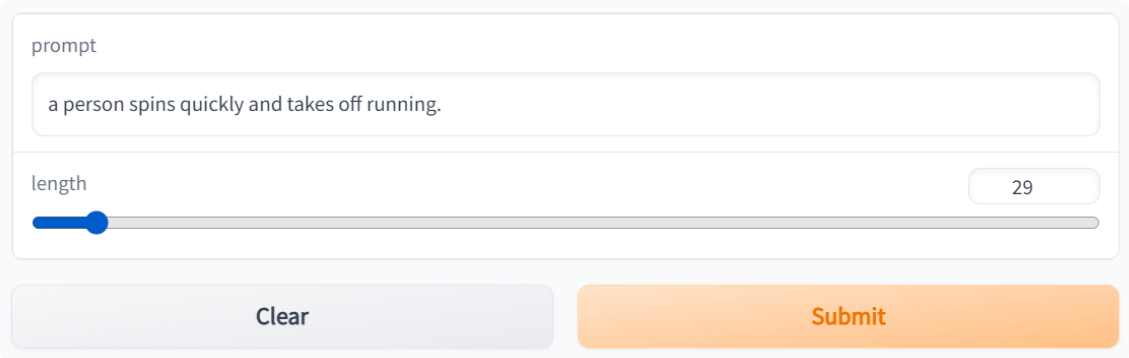
b) *Jumping* and *raising arms*

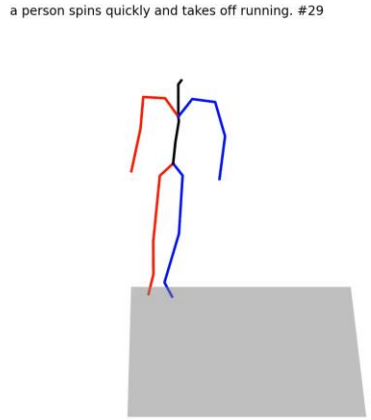
c) *Looking around* and then *calling on a phone with right hand*

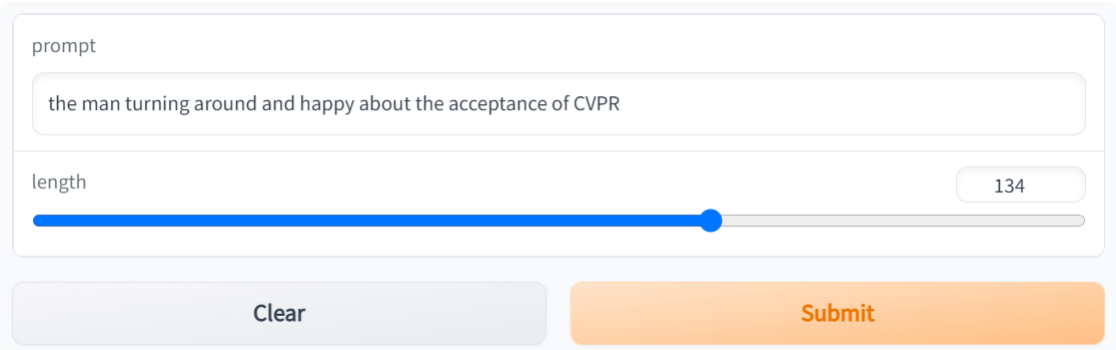
d) *Tying the shoe, standing up* and then *walking forward*

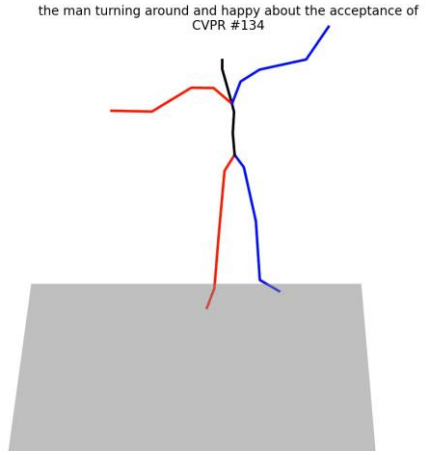

play the guitar
play the piano
play the violin

Examples


prompt
a person spins quickly and takes off running.
length
29
Clear
Submit


a person spins quickly and takes off running. #29


prompt
the man turning around and happy about the acceptance of CVPR
length
134
Clear
Submit


the man turning around and happy about the acceptance of CVPR #134

Online Demo

MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model

This is an interactive demo for MotionDiffuse. For more information, feel free to visit our project page(https://mingyuan-zhang.github.io/projects/MotionDiffuse.html).

prompt

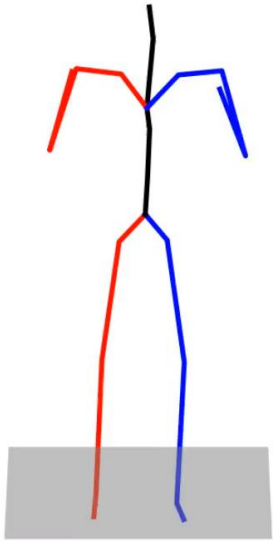
the man throws a punch with each hand.

length 58

Clear Submit

output

the man throws a punch with each hand. #58



2D Generation

3D Generation

Data

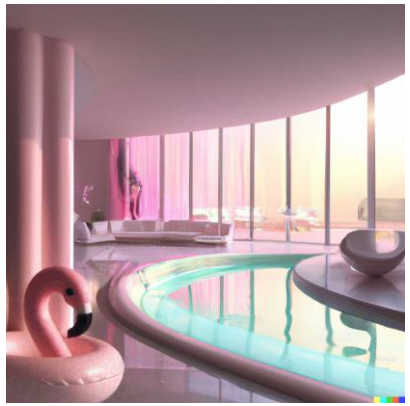
Software

Motion Generation

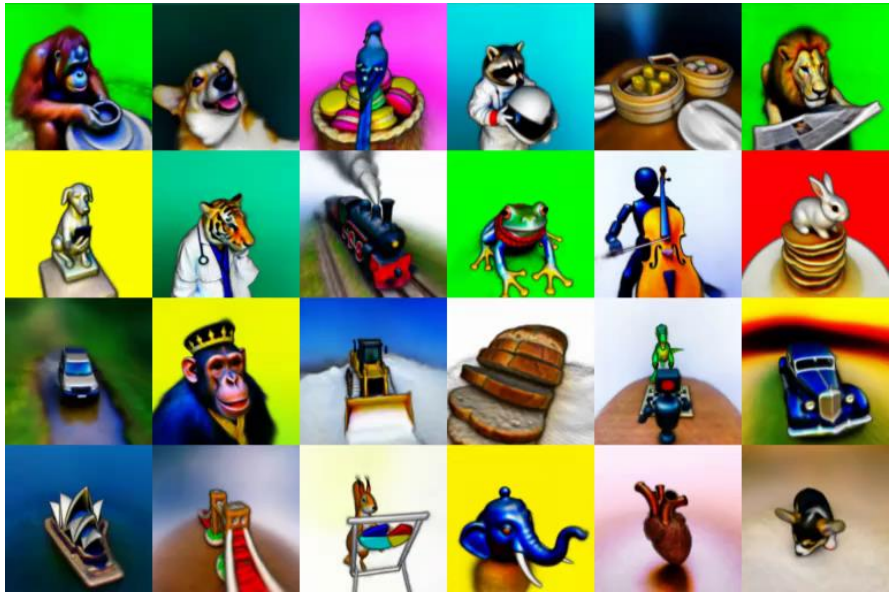
Scene Generation

Text-driven Content Generation

Imagen[1]



DALL·E[2]



DreamFusion[3]

[1] https://imagen.research.google/

[2] https://openai.com/dall-e-2/

[3] https://dreamfusion3d.github.io/

What about creating the environment?

The surrounding environment is also important to **an immersive VR experience.**

↓

Full field of view (360°) → Panorama

Realistic illuminations → HDR

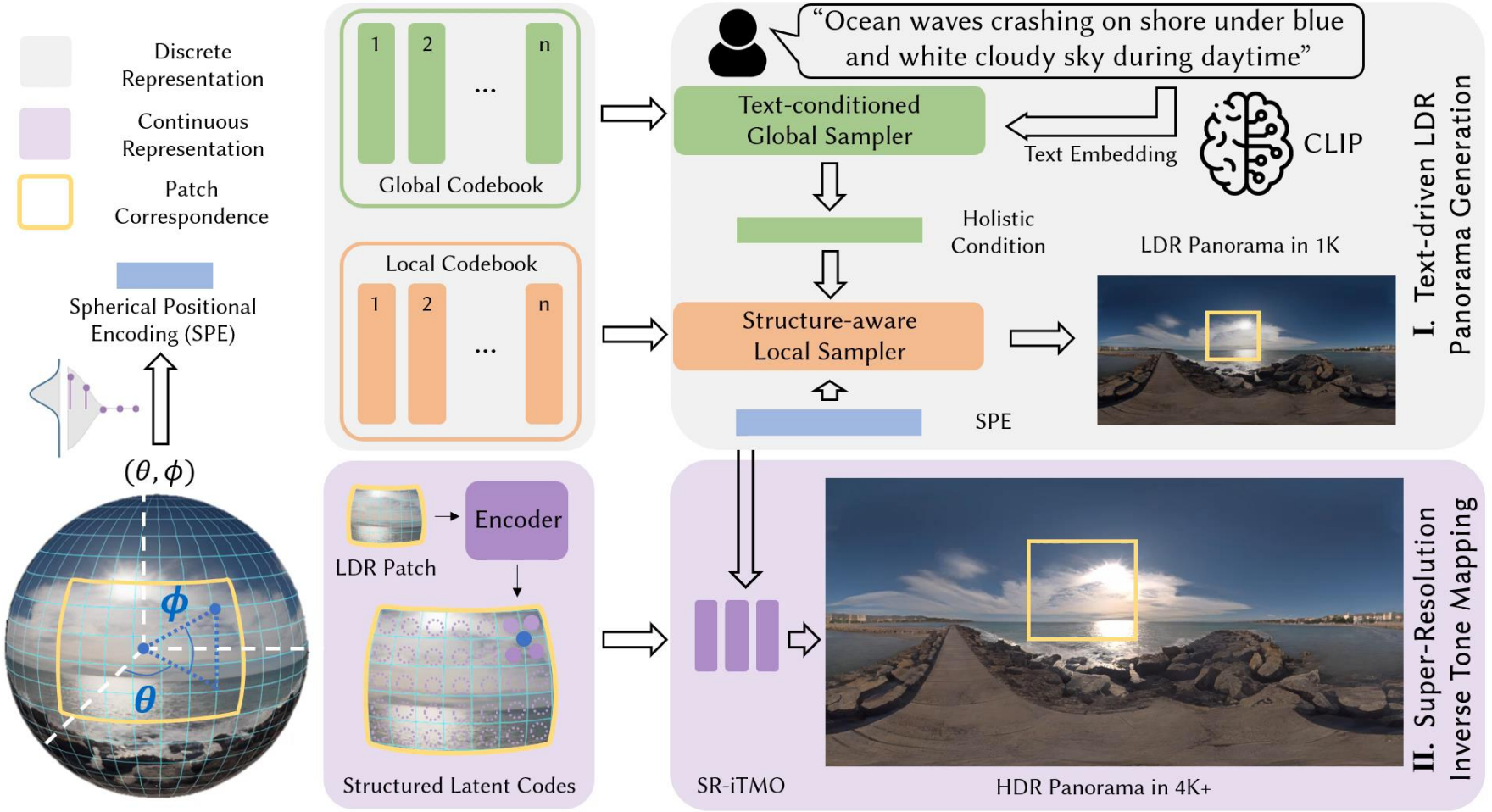
High-quality textures → 4K resolution

LDR – Low Dynamic Range, [0, 255]

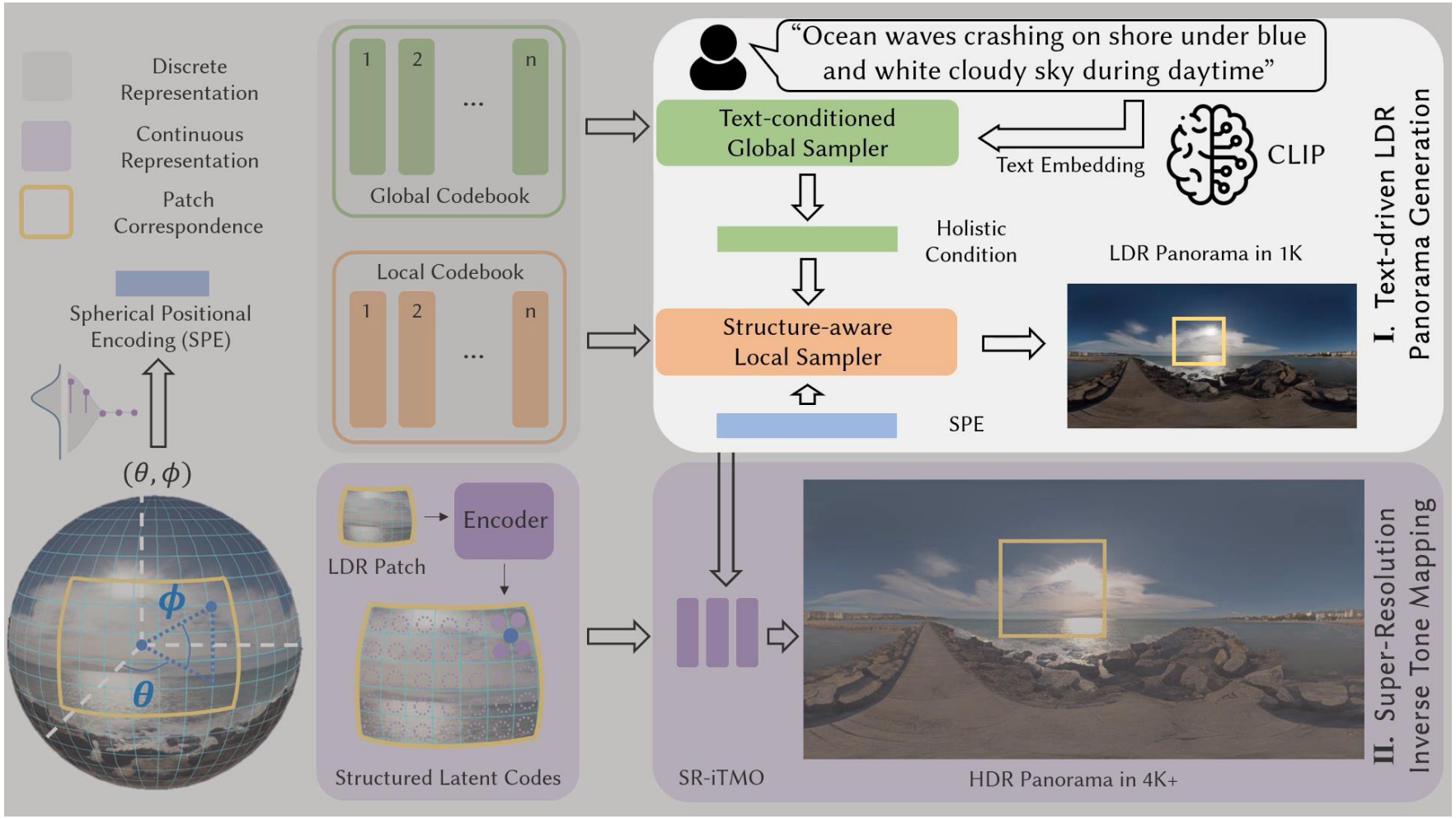
HDR – High Dynamic Range, [0, +∞]

Create the Surroundings Using Texts

Text2Light
An Overview



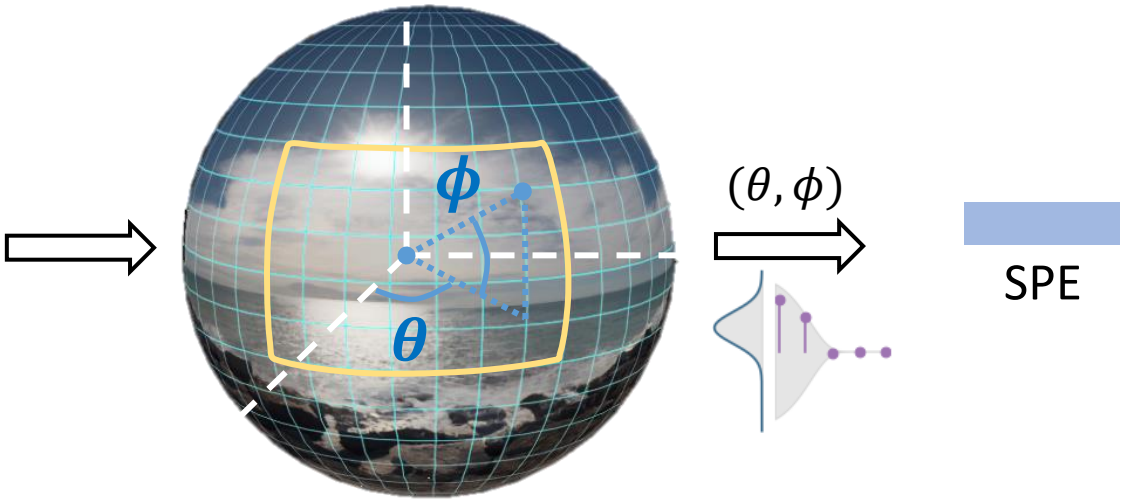
Text2Light
Stage I: Text-driven LDR Panorama Generation



Text2Light
Stage I: Structure-aware Local Sampler

Spherical Positional Encoding (SPE)

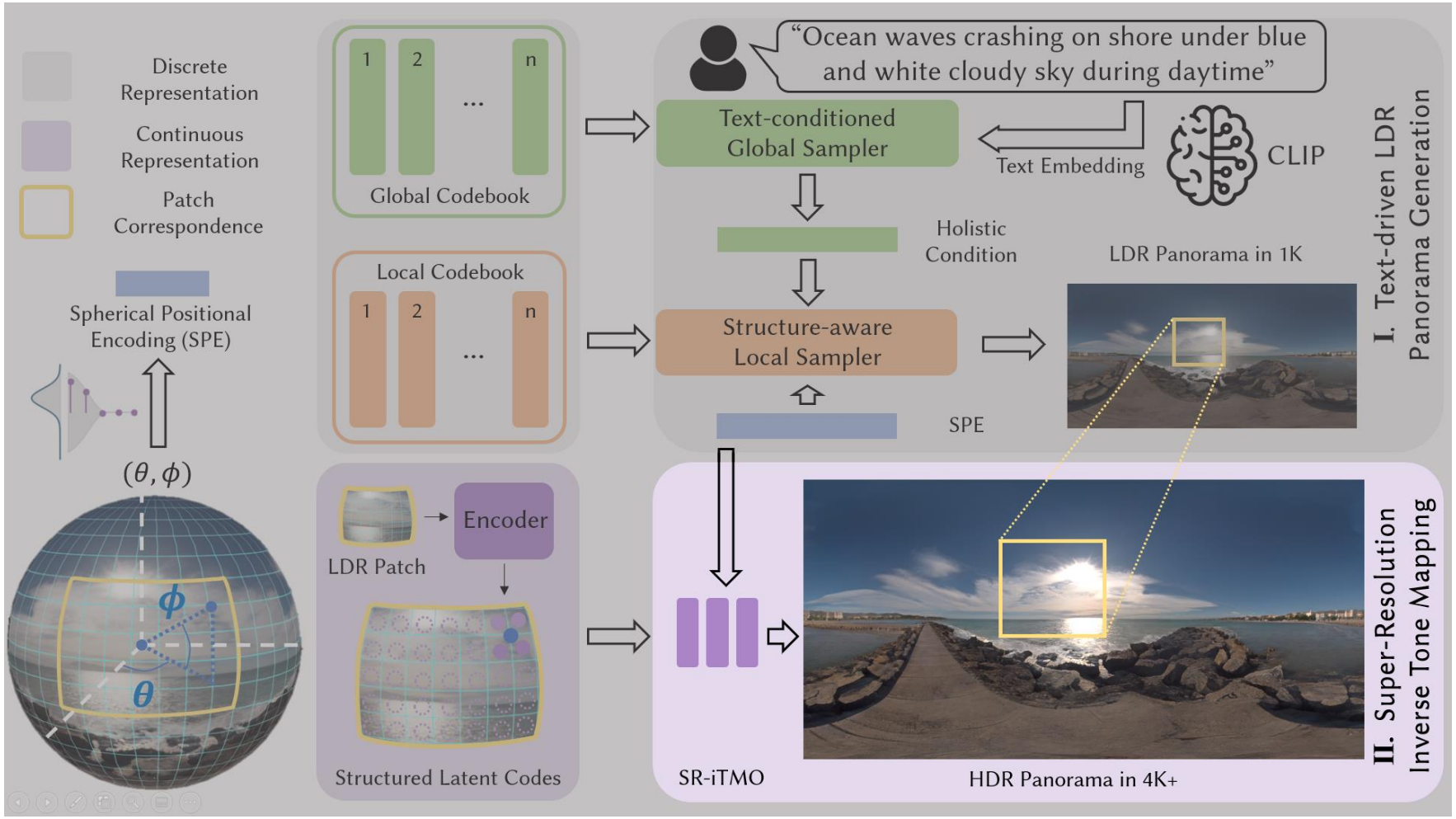
Raw HDR Panorama



Spherical Fields

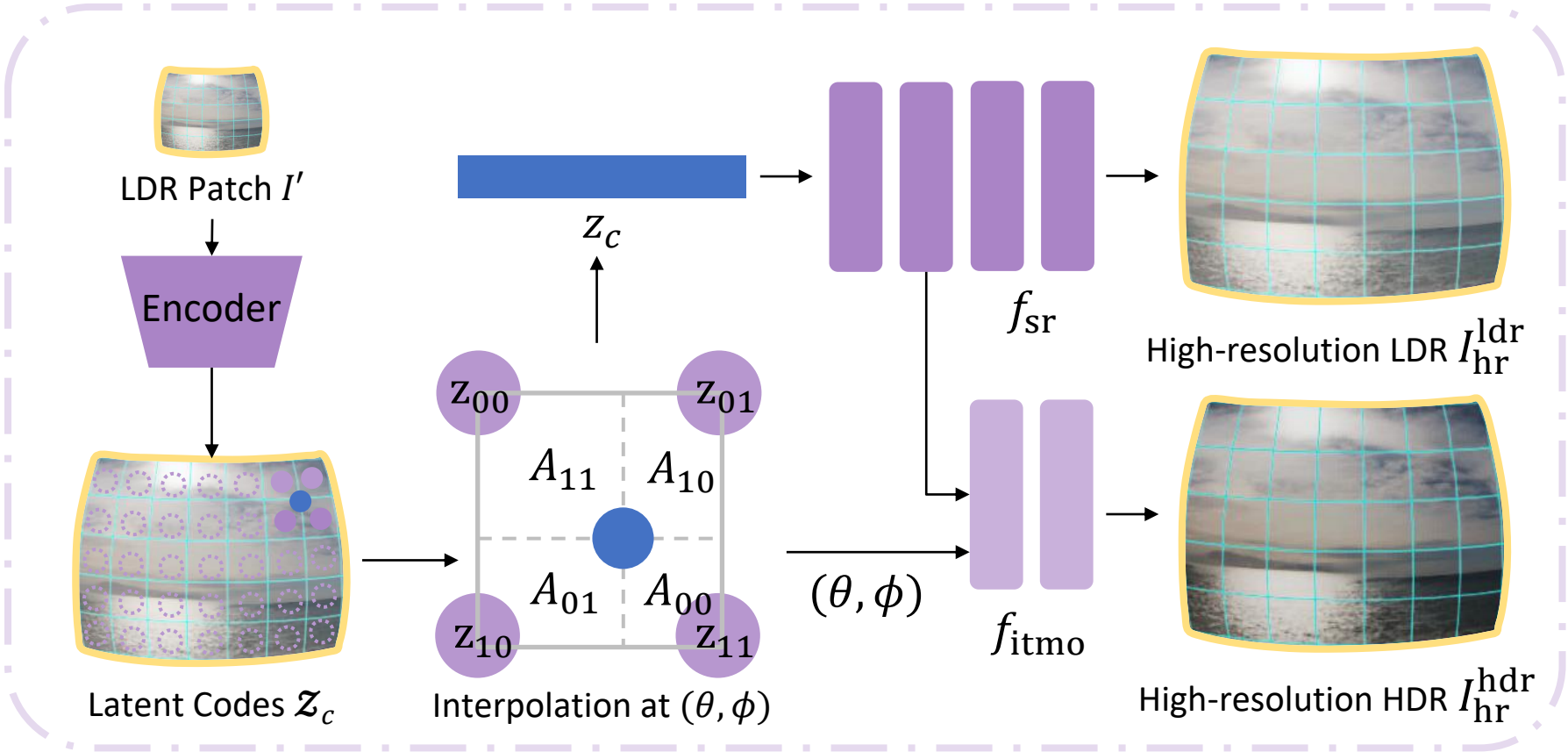
Fourier Features

Text2Light
Stage II: Super-Resolution Inverse Tonemapping



Text2Light
Stage II: SR-iTMO as two MLPs

Super-Resolution Inversed Tone Mapping Operator (SR-iTMO)



LDR – Low Dynamic Range
HDR – High Dynamic Range
LR – Low Resolution
HR – High Resolution

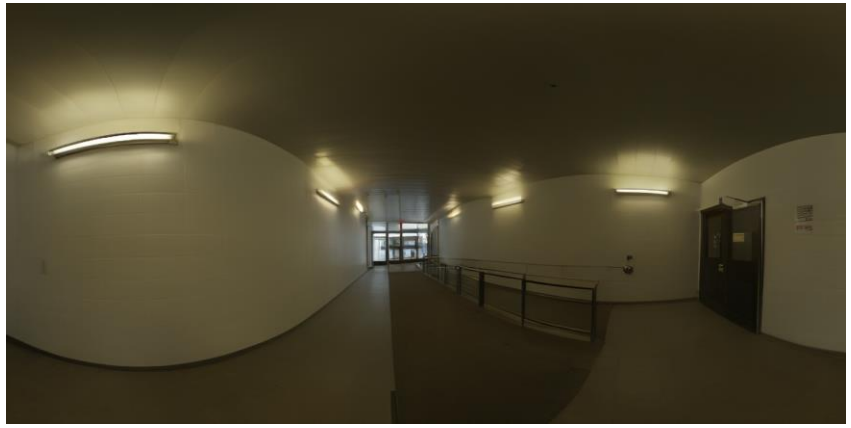
Text2Light Applications: UI

“white bed linen with white pillow”

“brown wooden floor with white wall”

“gray concrete pathway with wall signages”

“closeup photo of concrete stair surrounded by white painted wall”

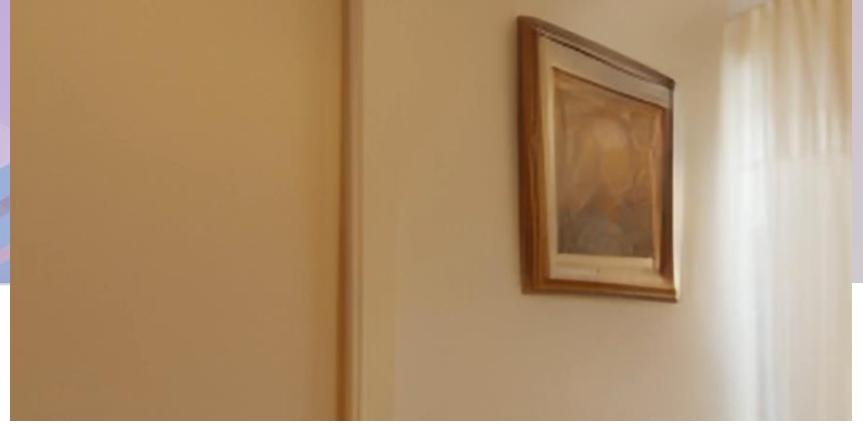


“blue and brown wooden counter”

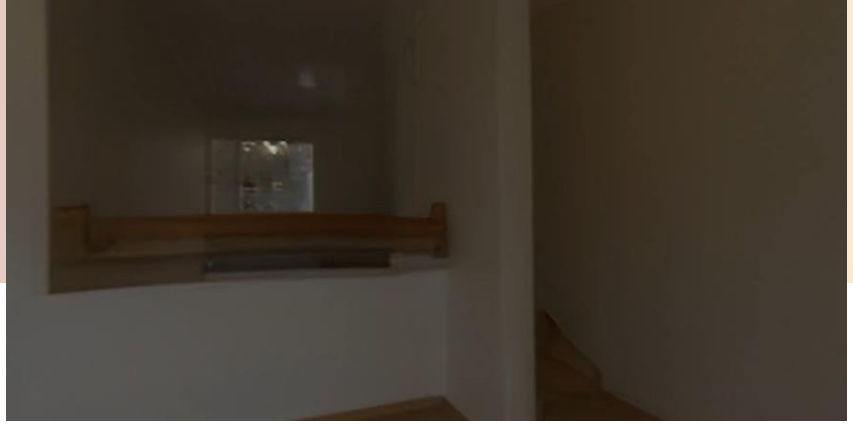
“empty parking lot during daytime”

Suzanne Monkey: glossy Shader balls: glass, diffuse, glossy, mixture of diffuse and glossy

“white bed linen with white pillow”

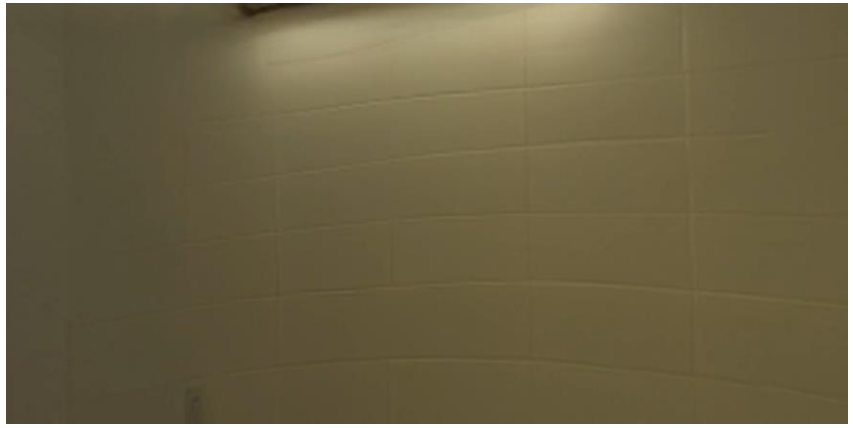


“brown wooden floor with white wall”

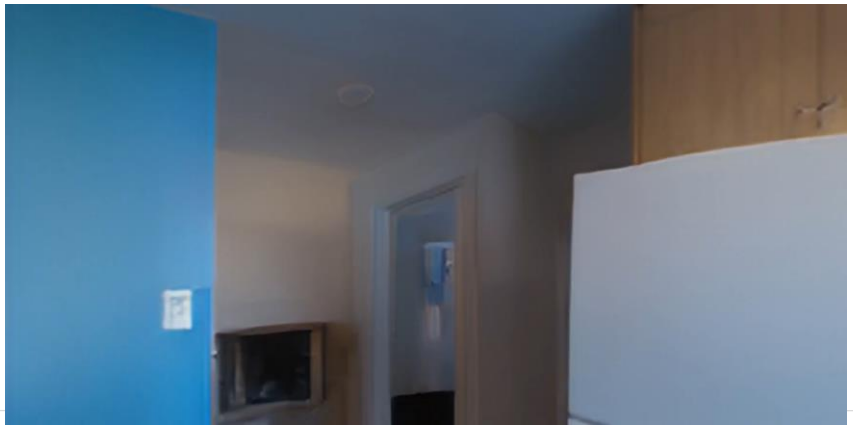


“gray concrete pathway with wall signages”

“closeup photo of concrete stair surrounded by white painted wall”



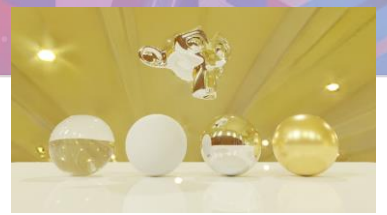
“blue and brown wooden counter”



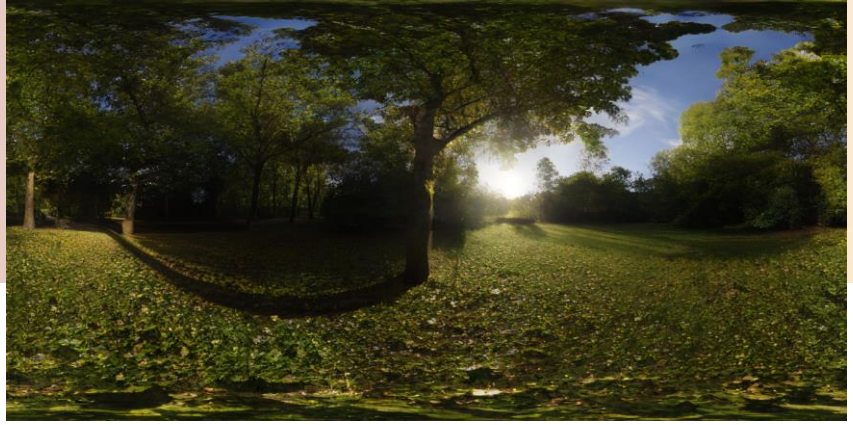
“empty parking lot during daytime”

Suzanne Monkey: glossy Shader balls: glass, diffuse, glossy, mixture of diffuse and glossy

“lined brown pew benches”

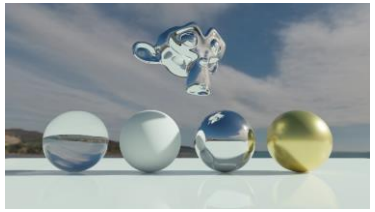


“Avenue, Trees, Path, Sunbeams, Sunrays”



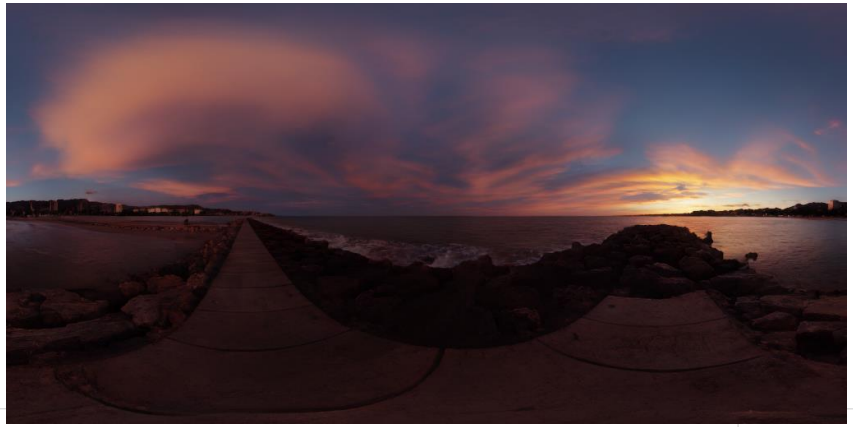
“photo of orange chairs”

“ocean waves crashing on shore under blue and white cloudy sky during daytime”



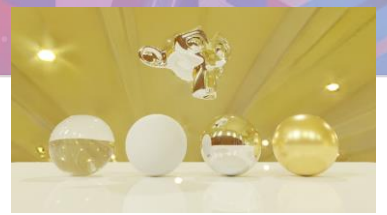
“road with falling leaves in between of trees”

“sunset by the ocean”

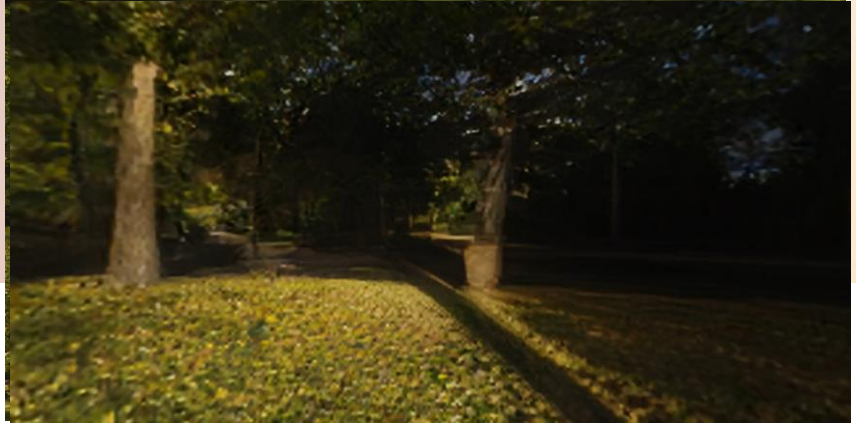


Suzanne Monkey: glossy Shader balls: glass, diffuse, glossy, mixture of diffuse and glossy

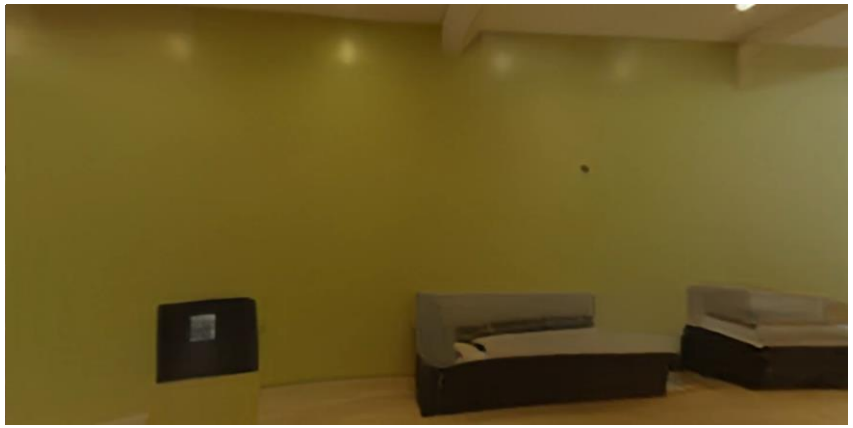
“lined brown pew benches”



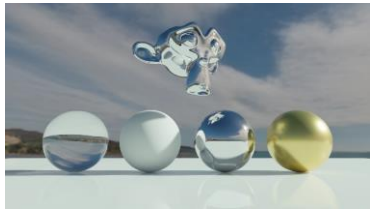
“Avenue, Trees, Path, Sunbeams, Sunrays”



“photo of orange chairs”

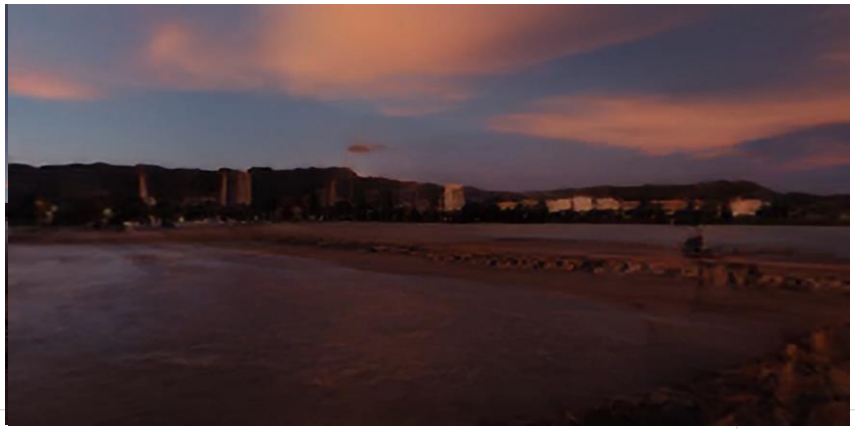


“ocean waves crashing on shore under blue and white cloudy sky during daytime”



“road with falling leaves in between of trees”

“sunset by the ocean”



Suzanne Monkey: glossy Shader balls: glass, diffuse, glossy, mixture of diffuse and glossy

Project Page

2D Generation

3D Generation

Data

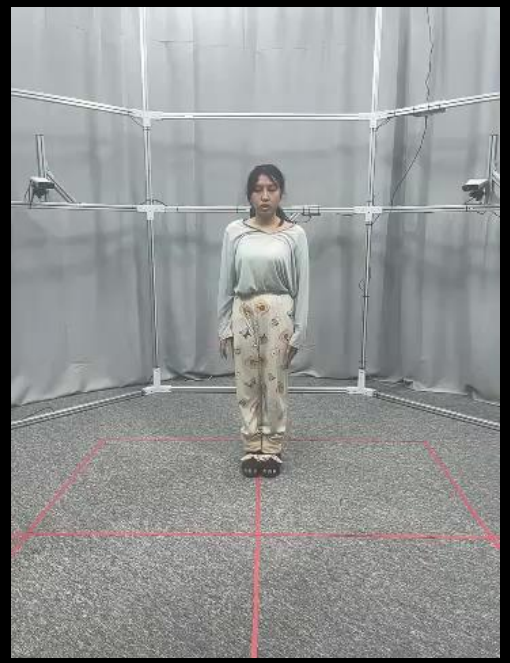
Software

Motion Generation

Scene Generation

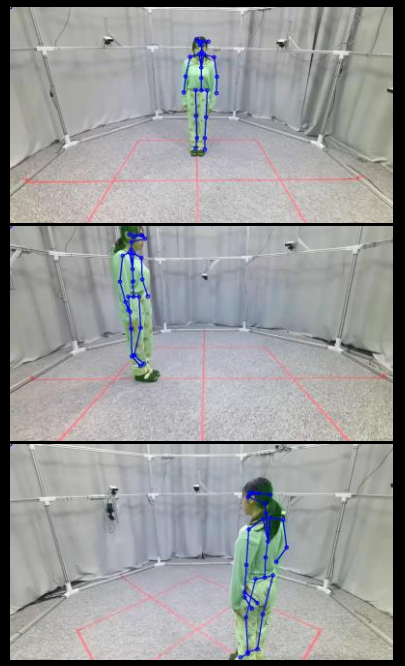
HuMMan MoCap System

Artec Eva

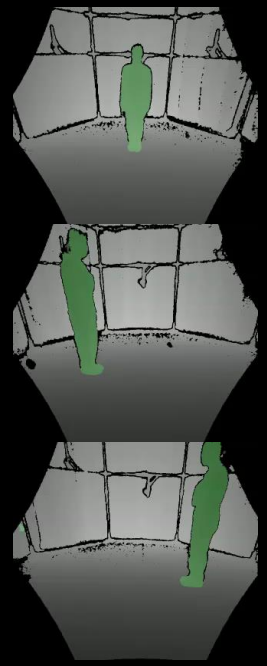


iPhone RGB

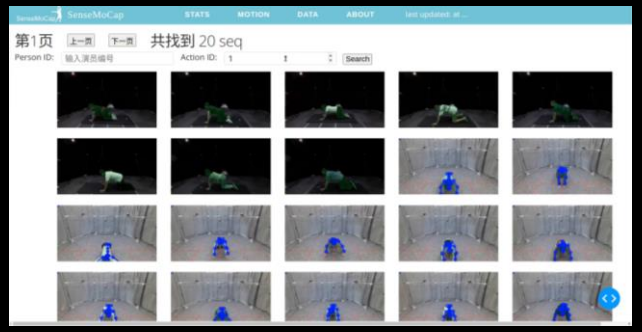
iPhone Depth



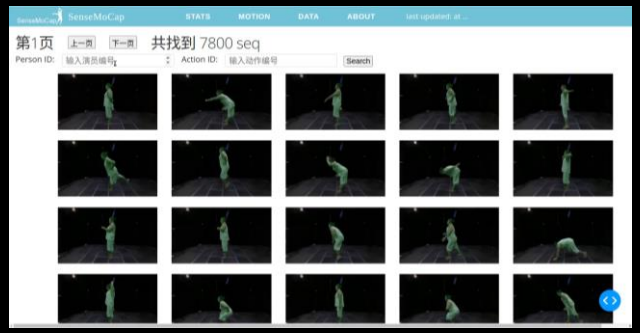
Kinect RGB



Kinect Depth



Search by Action



Search by Actor

Institute	#Camera	Mobile
Tencent	3	-
Tsinghua	6	-
ByteDance	6	-
NJU	9	-
CMU	10	-
Ours	**≥10**	✓

0.1mm
Accuracy

11
Cameras

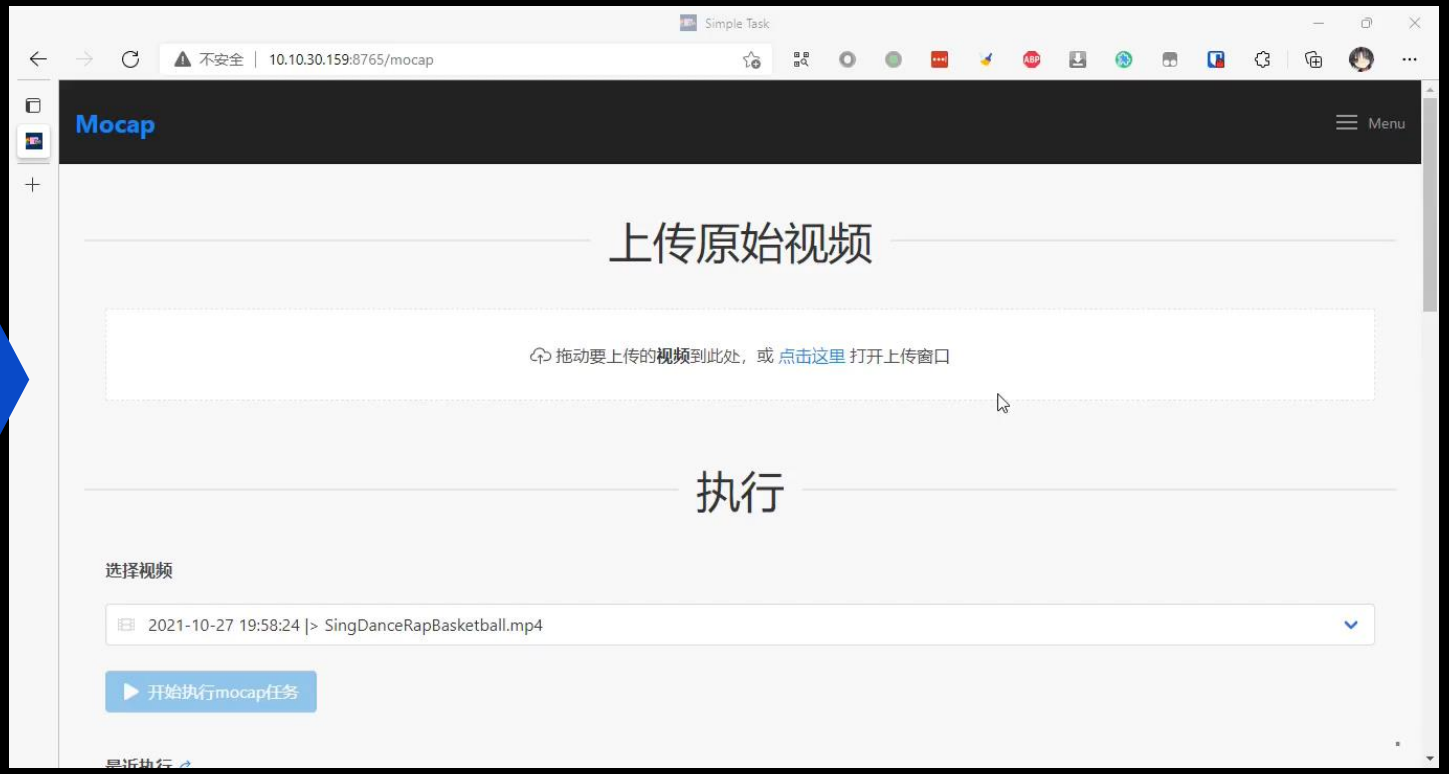
1G
Data / Sec

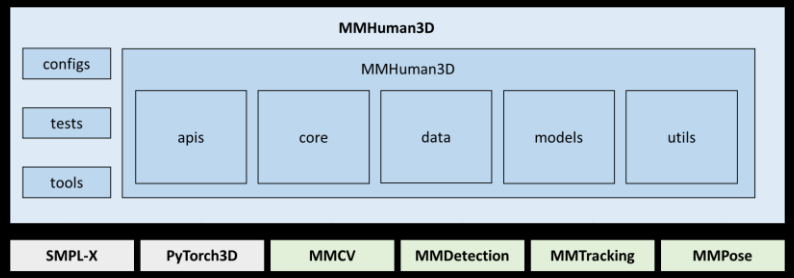
6
Actor / Day

MMHuman3D Software

Input Video

Online Motion Capture





3D Animation Production:

3 days -> 30 min

2D Generation

3D Generation

Thank you!

Motion Generation

Scene Generation