SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

商汤 sensetime

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

THE PREMIER CONFERENCE & EXHIBITION ON
COMPUTER GRAPHICS & INTERACTIVE TECHNIQUES

Fangzhou Hong[1*]

Mingyuan Zhang[1*]

Liang Pan[1]

Zhongang Cai[1,2,3]

Lei Yang[2]

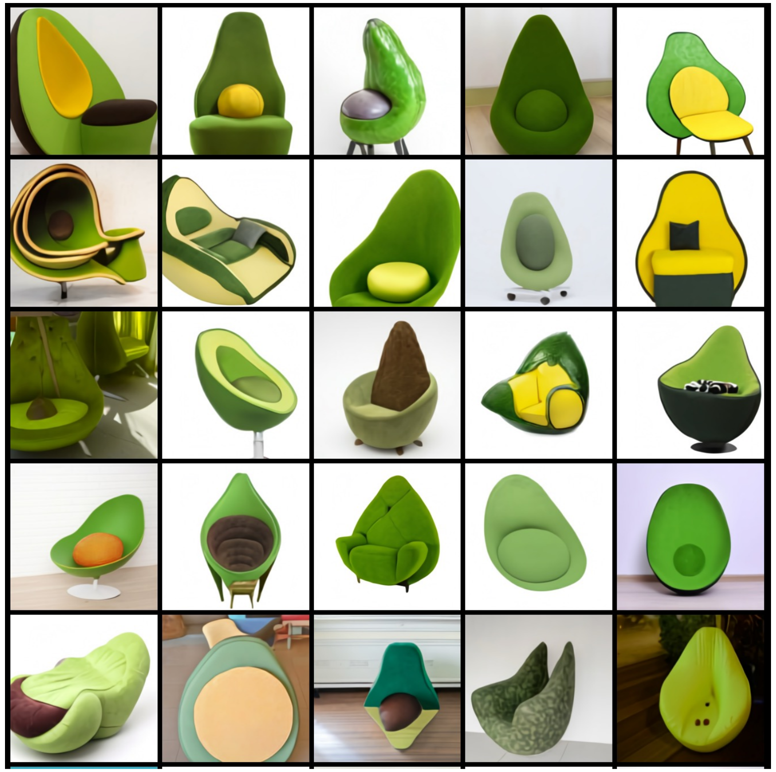
Ziwei Liu[1+]

[1]S–Lab Nanyang Technological University [2]SenseTime Research [3]Shanghai AI Laboratory
[*]Equal Contribution [+]Corresponding Author

AvatarCLIP

ZERO-SHOT TEXT-DRIVEN GENERATION AND ANIMATION OF 3D AVATARS

TEXT-DRIVEN IMAGE GENERATION



DALL·E [1]

DALL·E 2 [2]

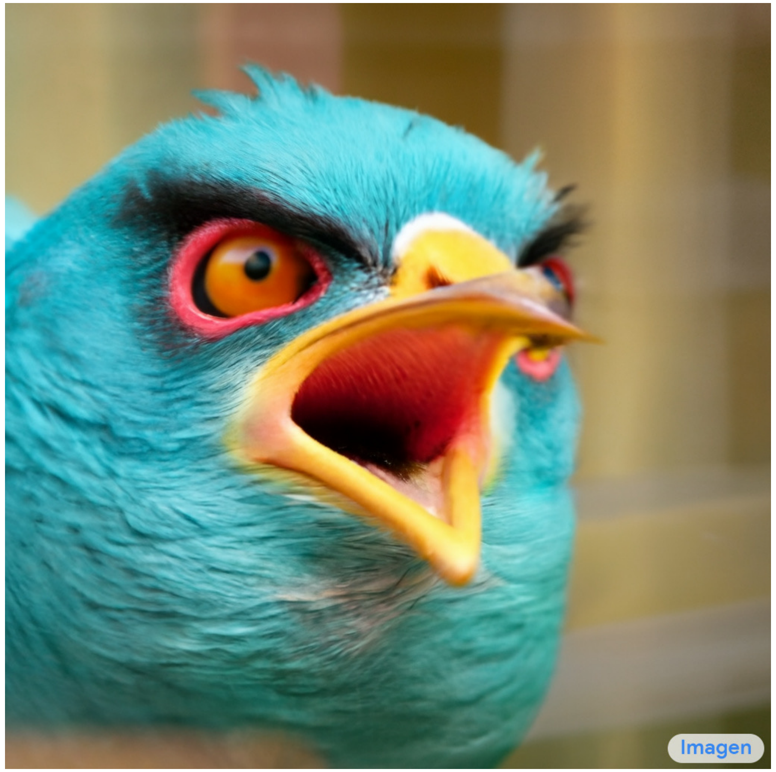
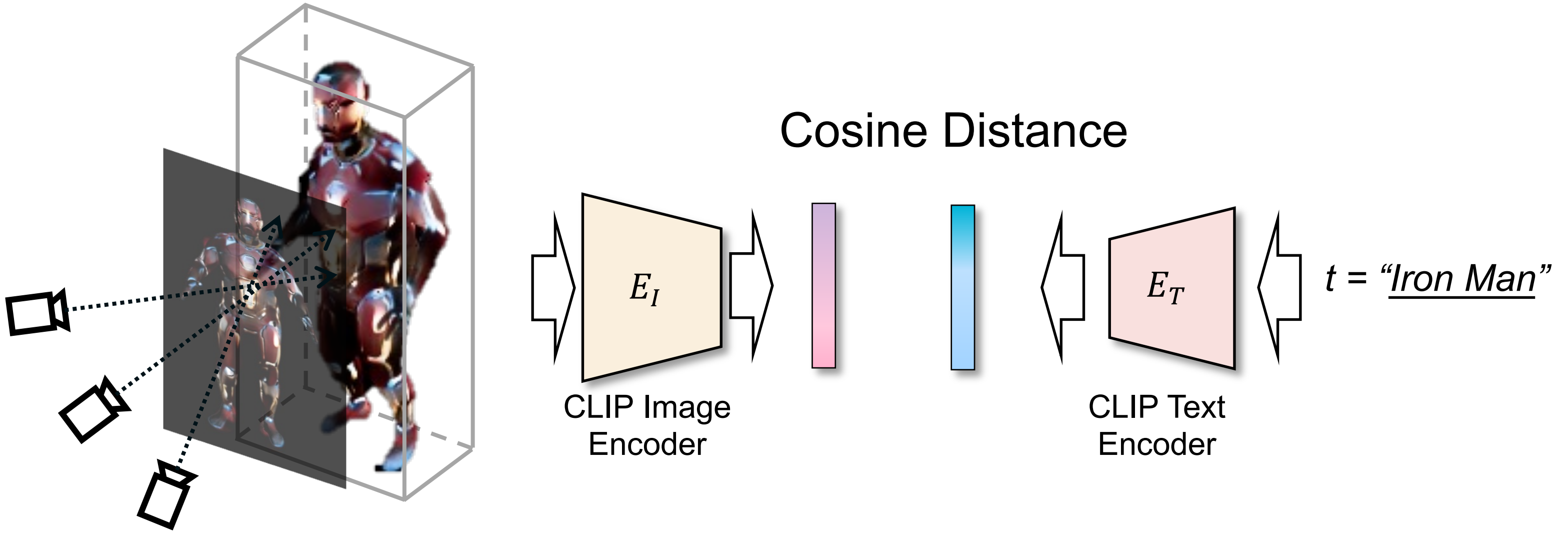


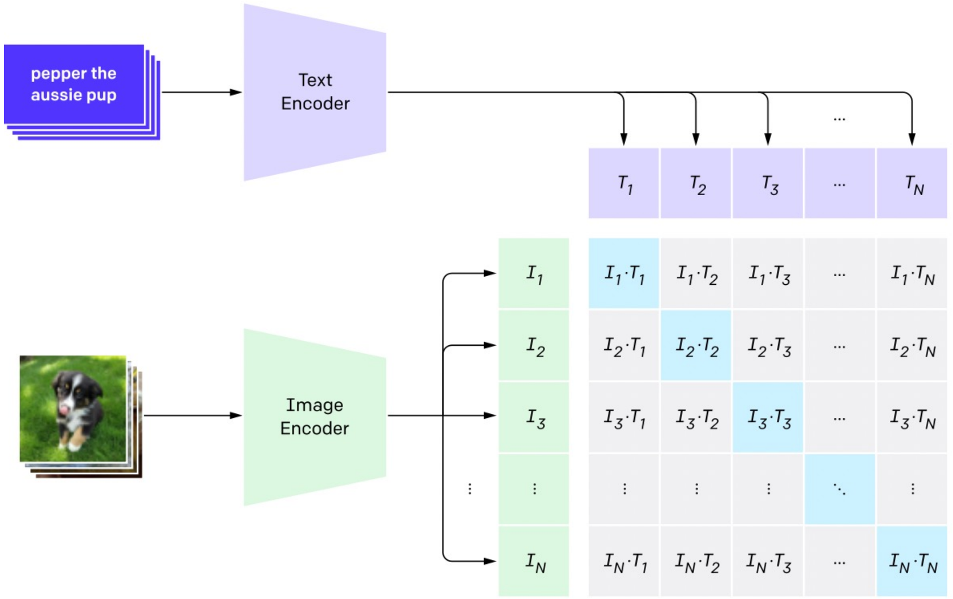
Imagen [3]

[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





	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

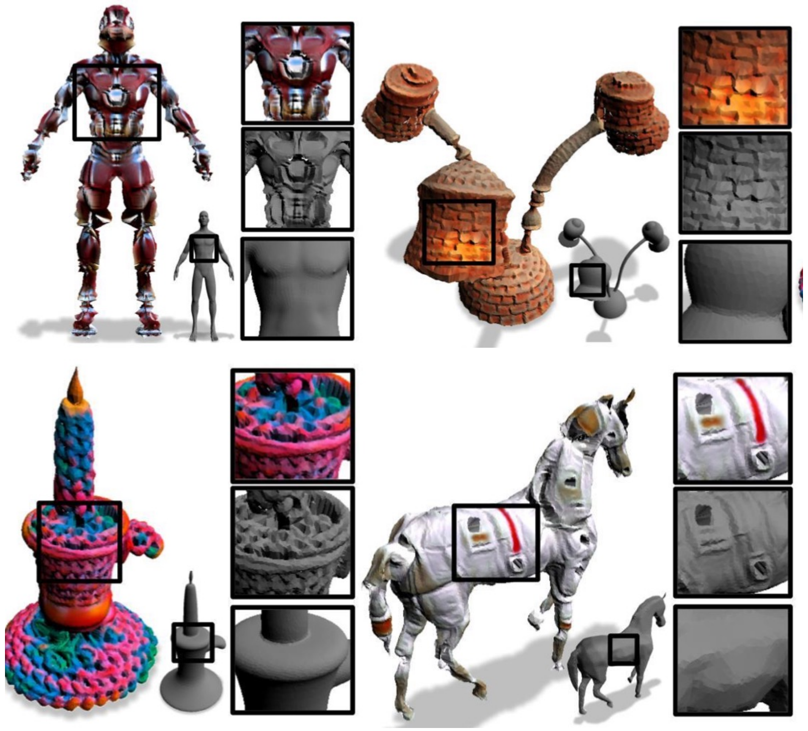
a) Differentiable Rendering

b) Optimization guided by CLIP

TEXT-DRIVEN 3D GENERATION

CLIP + DIFFERENTIABLE RENDERING

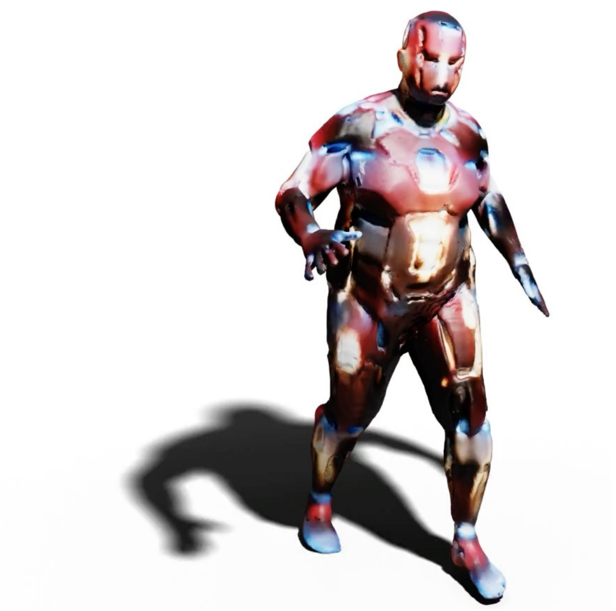
Dream Field [1]



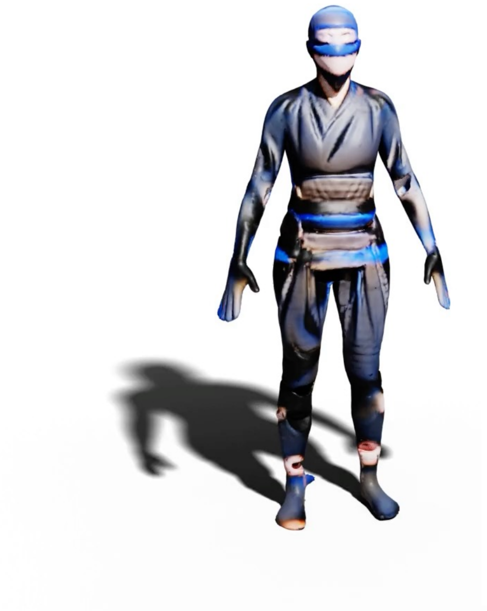
Text2Mesh [2]

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

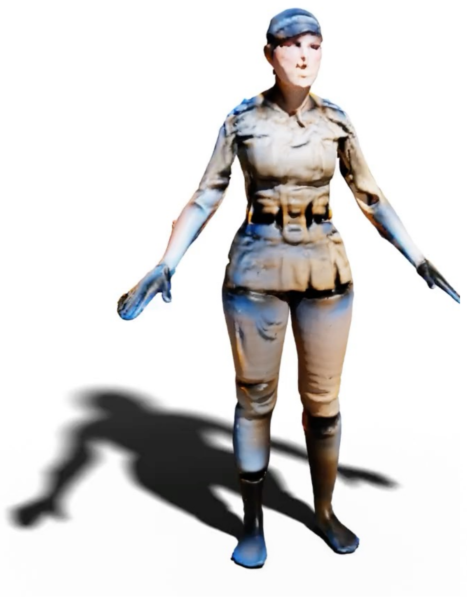
WHAT ABOUT TEXT-DRIVEN AVATAR GENERATION => NOW WE HAVE AVATARCLIP



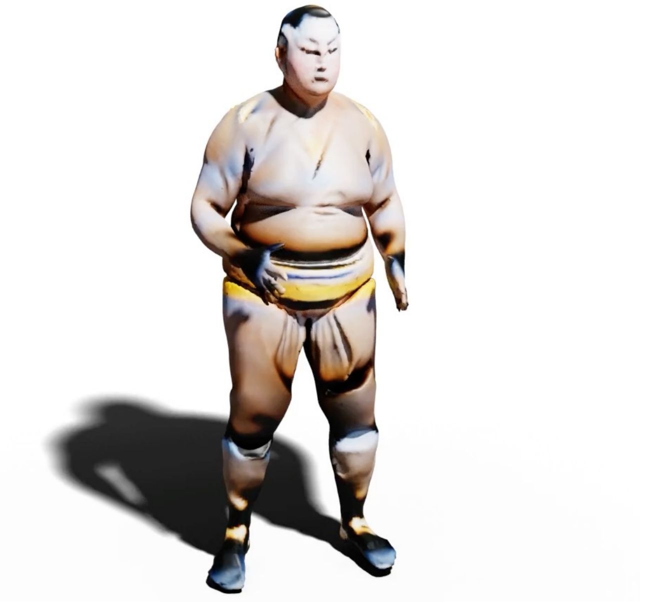
I want to generate a tall and fat Iron Man that is running.



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.



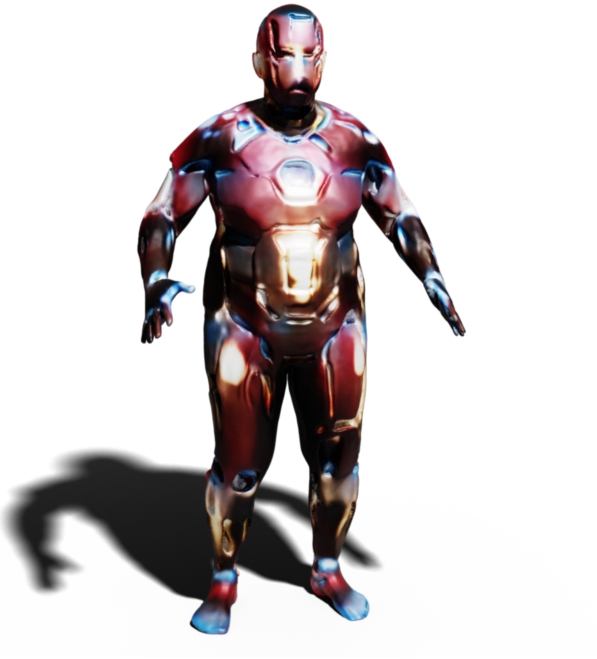
I want to generate an overweight sumo wrestler that is sitting.

AVATARCLIP: HOW IT WORKS

A) STATIC AVATAR GENERATION

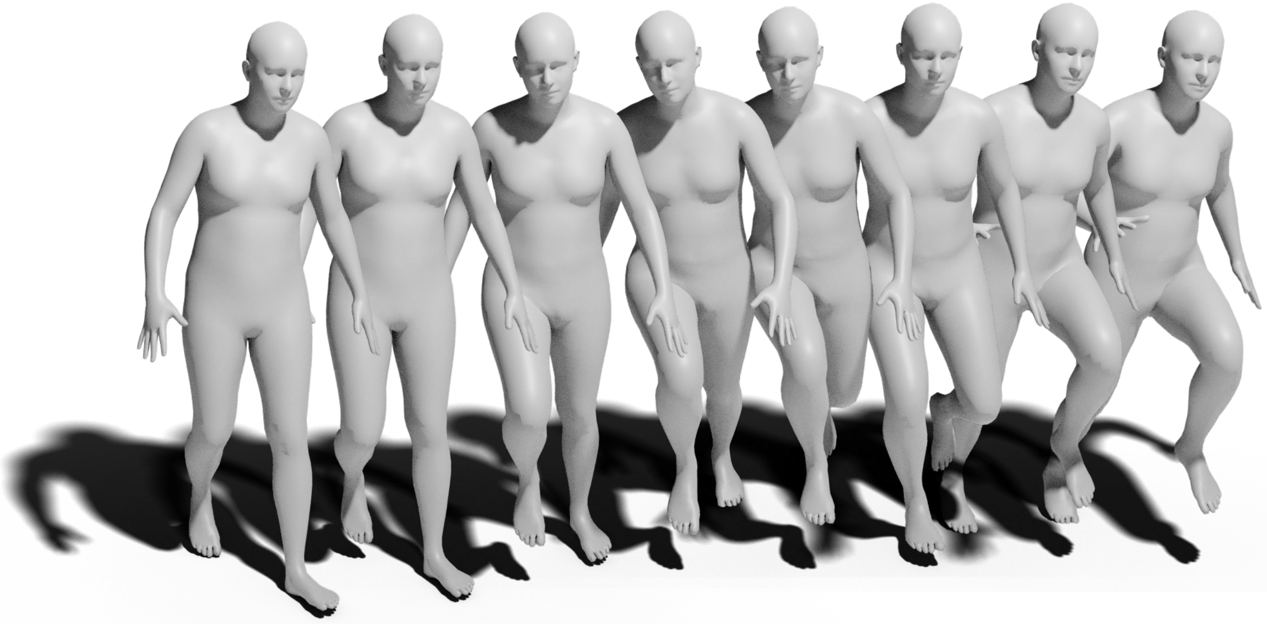
Shape Description: *“a tall and fat man”*

Appearance Description: *“Iron Man”*

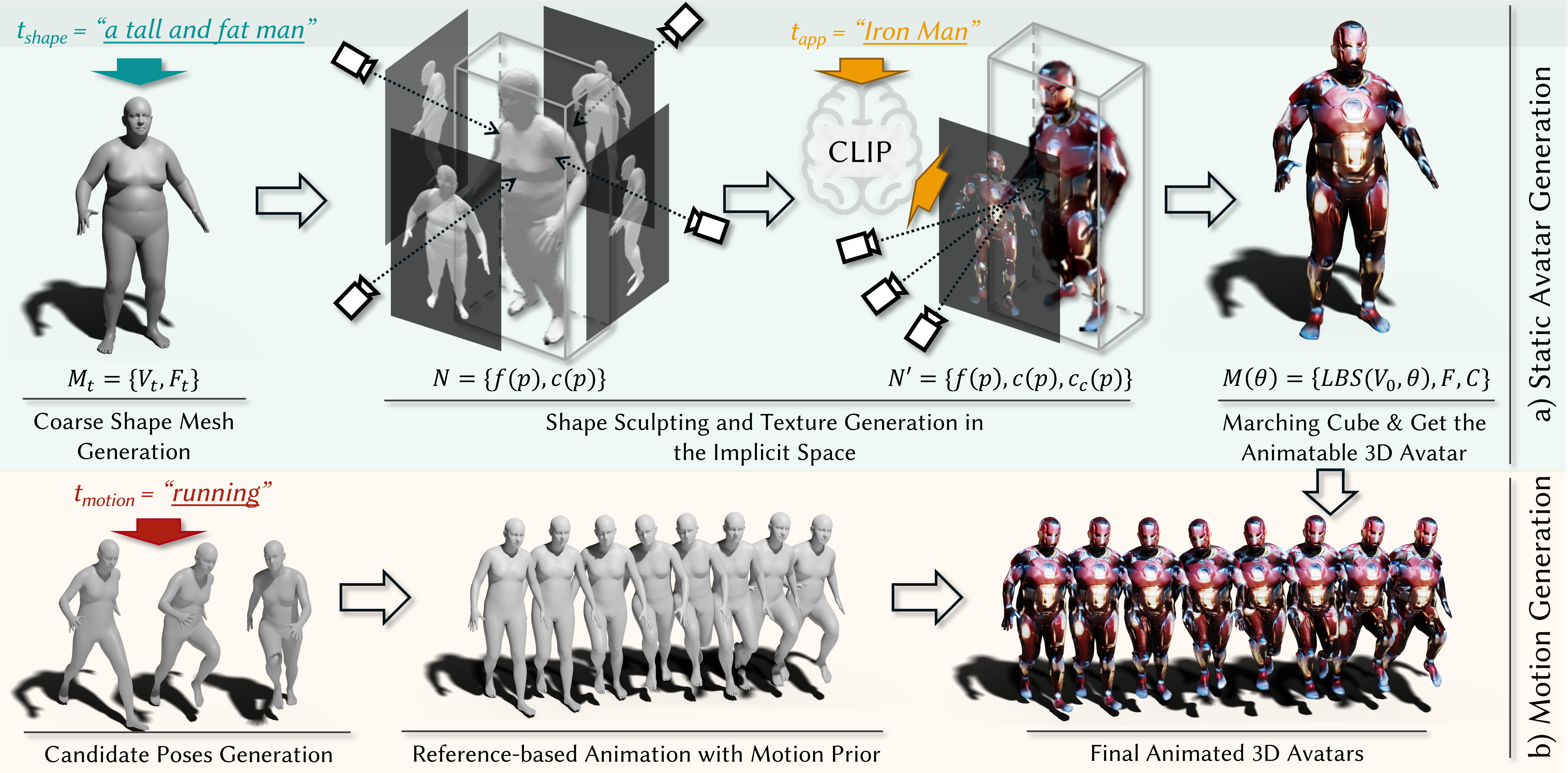


B) MOTION GENERATION

Motion Description: *“running”*



AVATARCLIP: DETAILED PIPELINE



t_{shape} = *"a tall and fat man"*

t_{app} = *"Iron Man"*

$M_t = \{V_t, F_t\}$

Coarse Shape Mesh Generation

$N = \{f(p), c(p)\}$

$N' = \{f(p), c(p), c_c(p)\}$

Shape Sculpting and Texture Generation in the Implicit Space

$M(\theta) = \{LBS(V_0, \theta), F, C\}$

Marching Cube & Get the Animatable 3D Avatar

a) Static Avatar Generation

t_{motion} = *"running"*

Candidate Poses Generation

Reference-based Animation with Motion Prior

Final Animated 3D Avatars

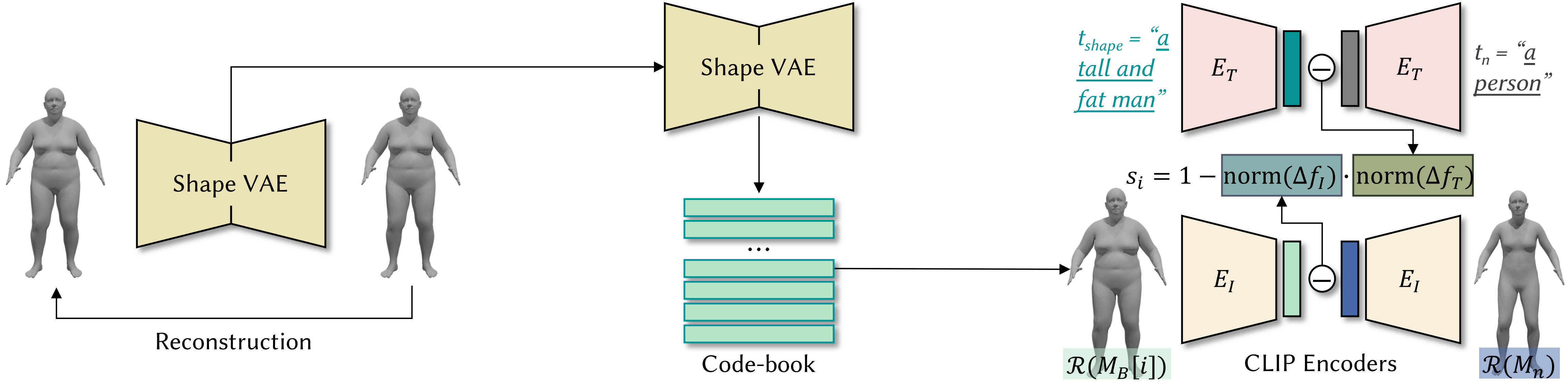
b) Motion Generation

AVATARCLIP: COARSE SHAPE GENERATION

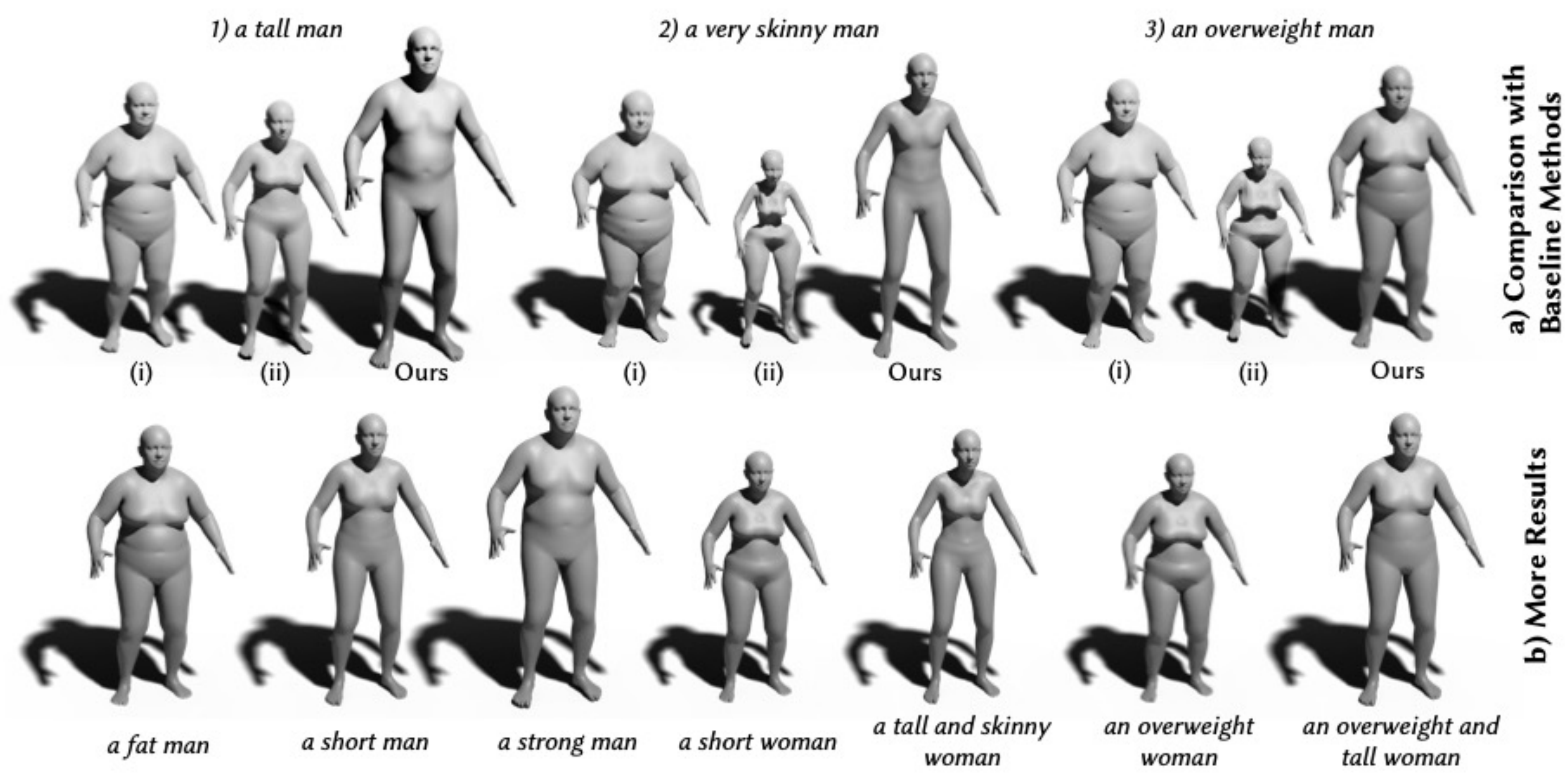
A) SMPL SHAPE VAE TRAINING

B) CODE-BOOK CONSTRUCTION

C) CLIP-GUIDED CODE-BOOK QUERY



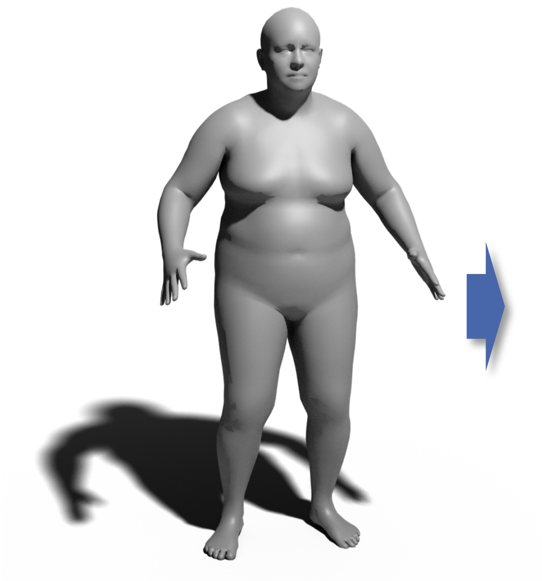
SHAPE GENERATION RESULTS & COMPARISON



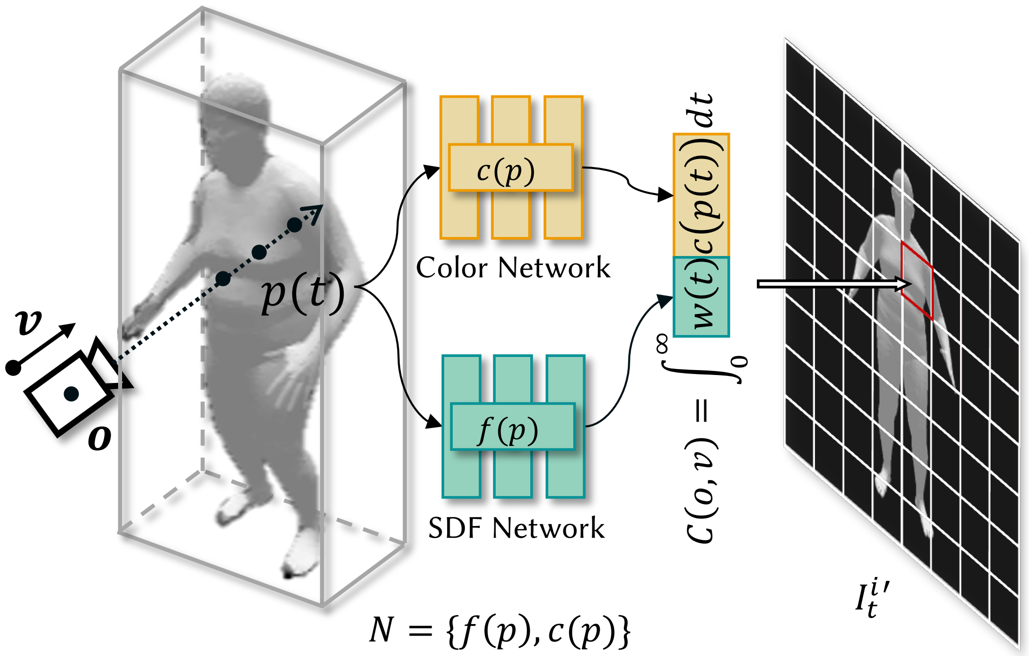
(i) Direct optimization on SMPL parameter beta

(ii) Direct optimization on shape VAE latent code

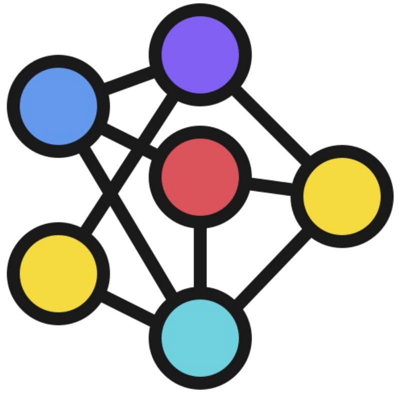
AVATARCLIP: TO THE IMPLICIT SPACE



Mesh

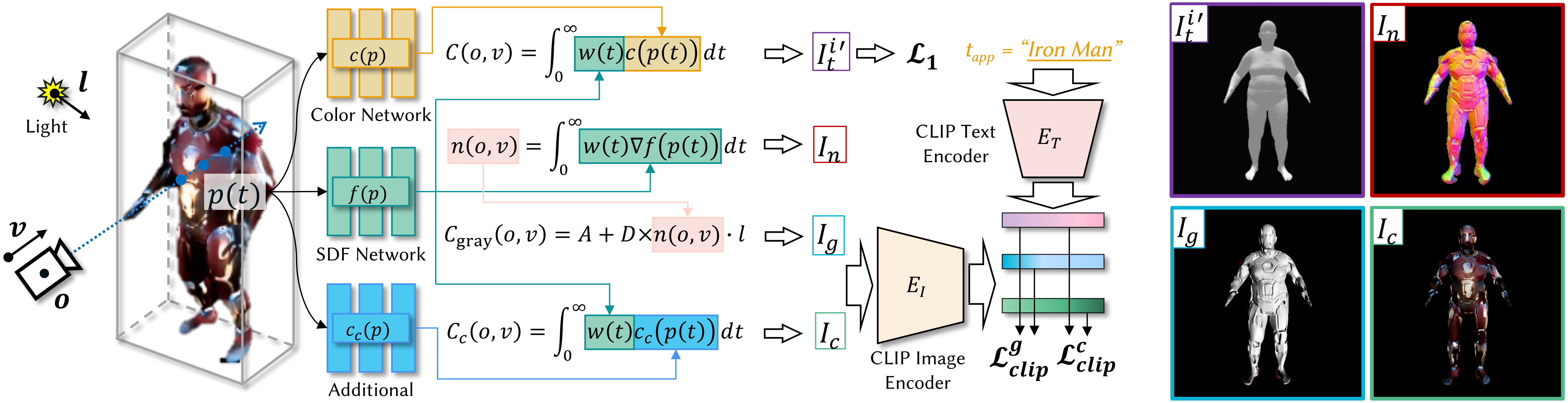


NeuS



Implicit Function

AVATARCLIP: SHAPE SCULPTING AND TEXTURE GENERATION



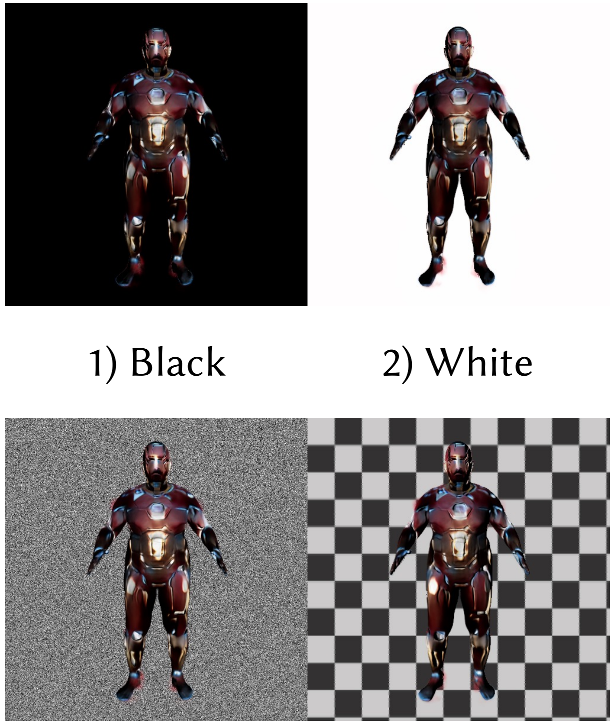
a) Rendering the Implicit 3D Avatar $N' = \{f(p), c(p), c_c(p)\}$

b) Optimization

Examples of Intermediate Results

AVATARCLIP: OPTIMIZATION PROCESS

A) RANDOM BACKGROUND AUGMENTATION



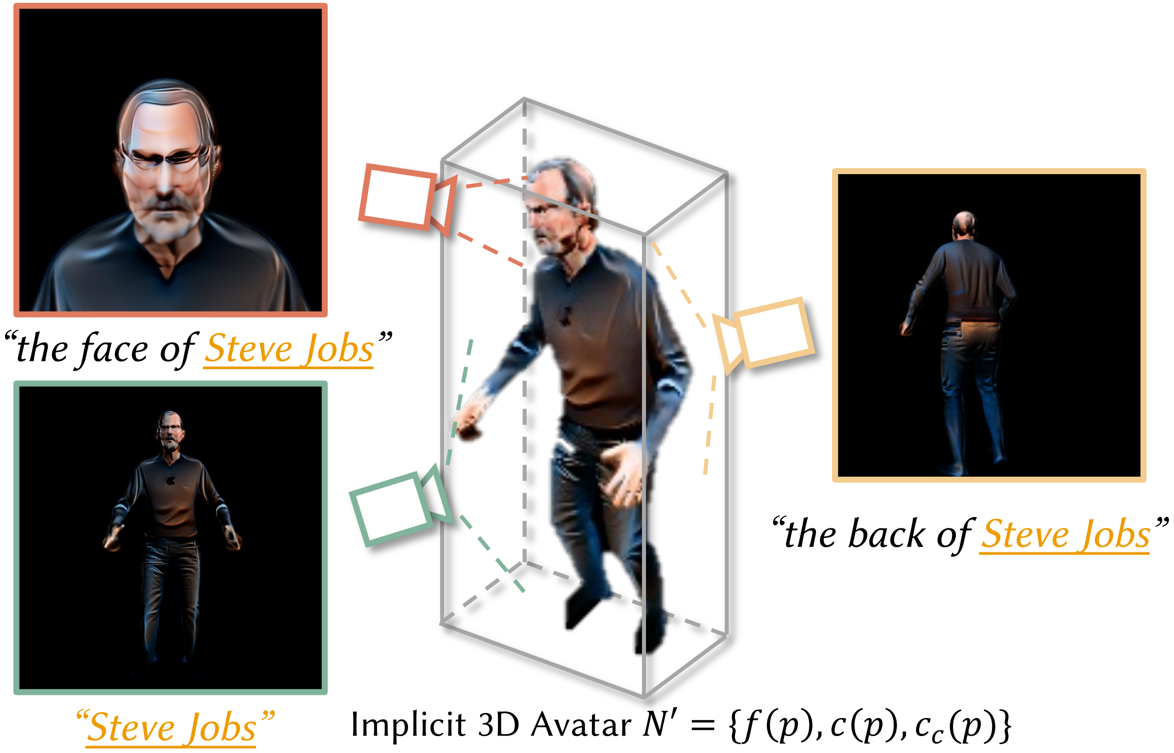
1) Black

2) White

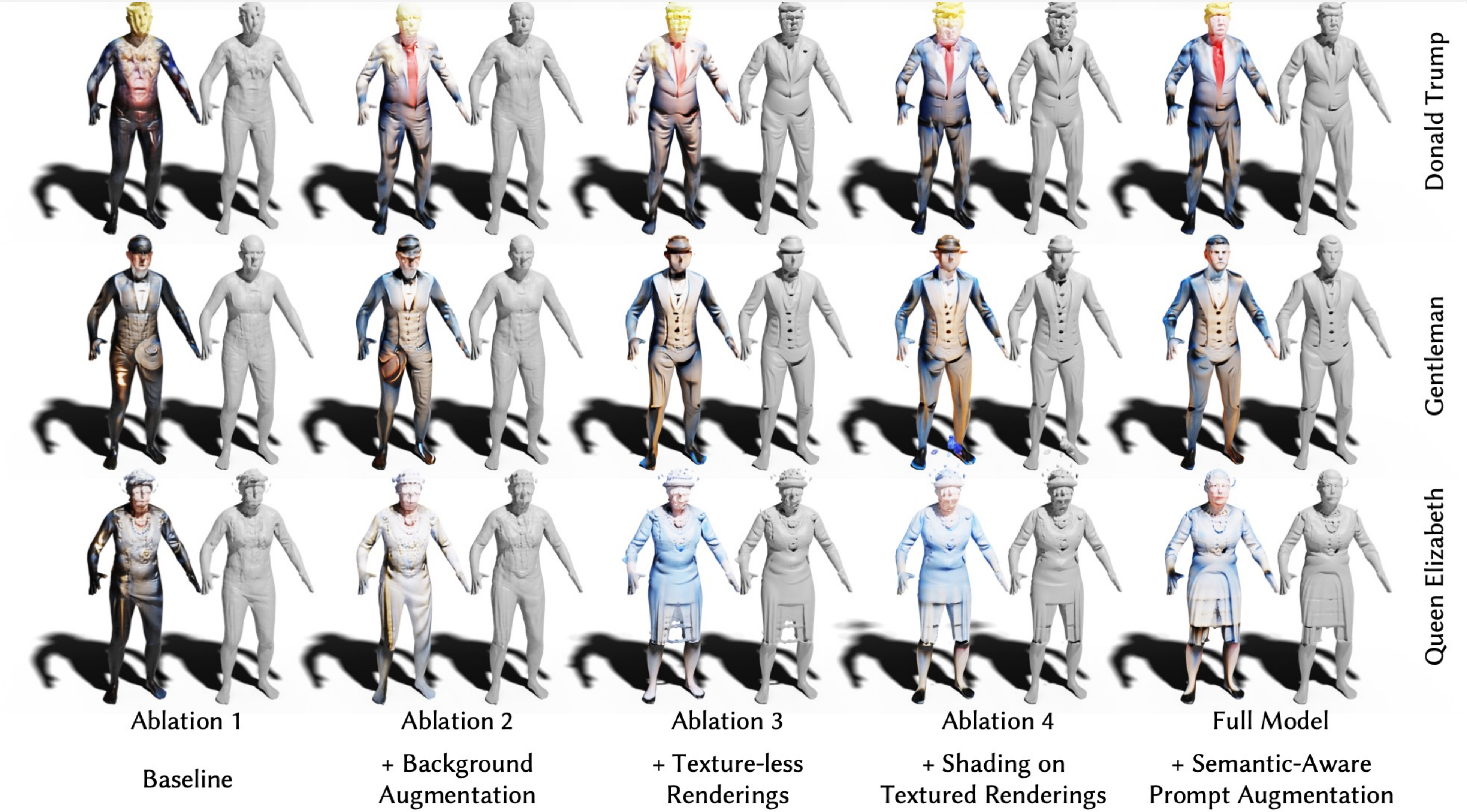
3) Gaussian Noise

4) Chess Board

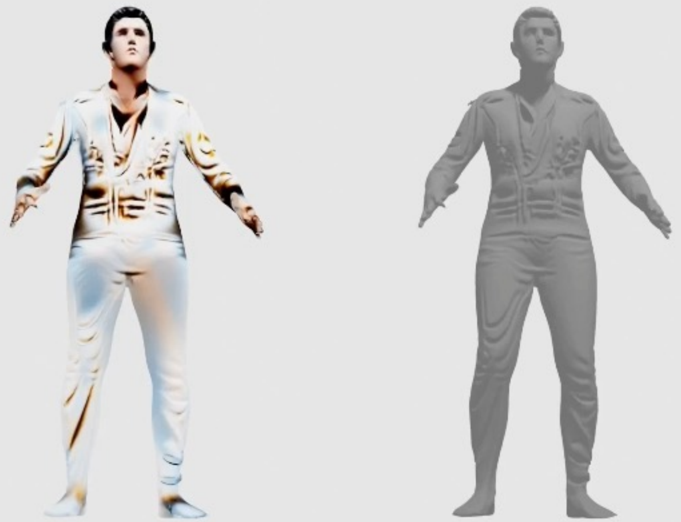
B) SEMANTIC-AWARE PROMPT AUGMENTATION



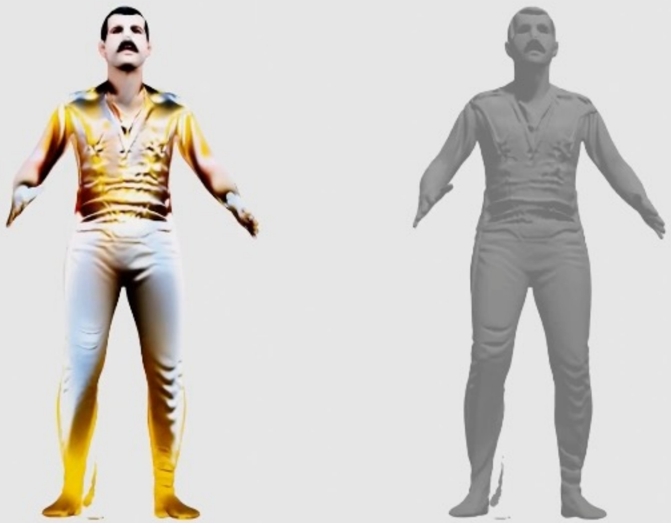
AVATAR GENERATION ABLATION


Donald Trump
Gentleman
Queen Elizabeth
Ablation 1
Baseline
Ablation 2
+ Background Augmentation
Ablation 3
+ Texture-less Renderings
Ablation 4
+ Shading on Textured Renderings
Full Model
+ Semantic-Aware Prompt Augmentation

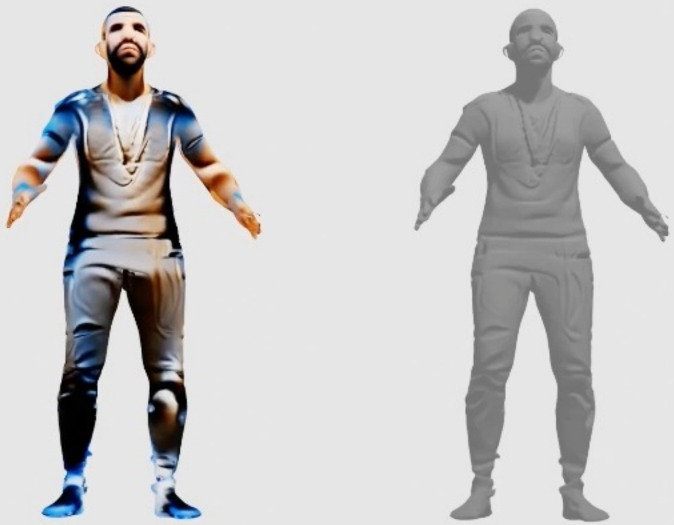
AVATAR GENERATION RESULTS



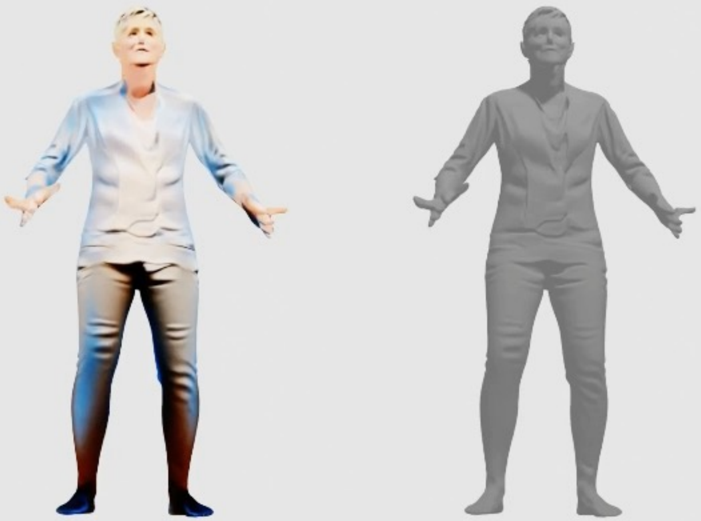
Elvis Presley



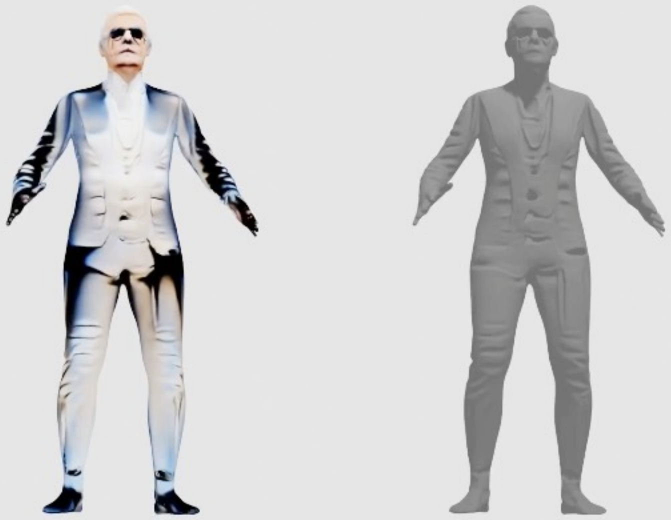
Freddie Mercury



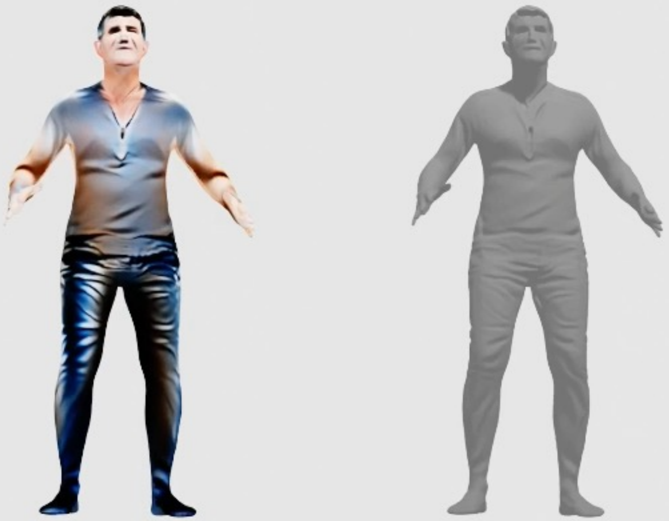
Drake



Ellen DeGeneres

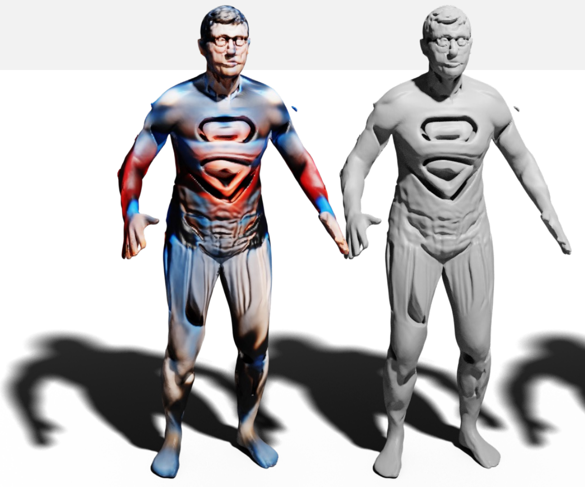


Karl Lagerfeld

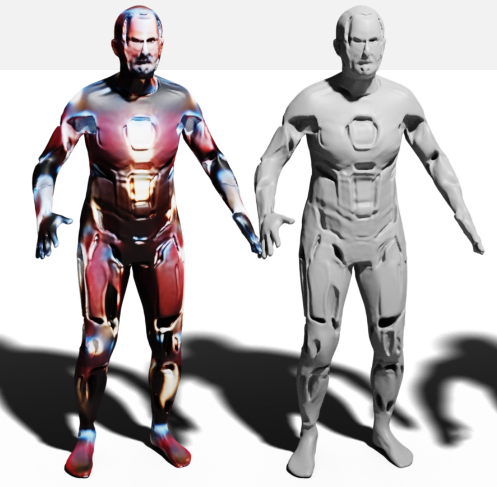


Simon Cowell

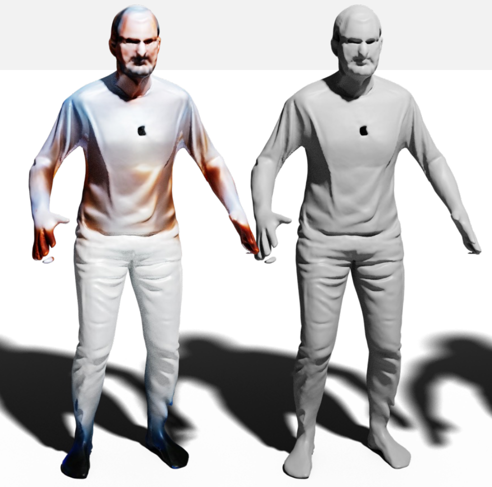
CONTROLLING & CONCEPT MIXING ABILITIES



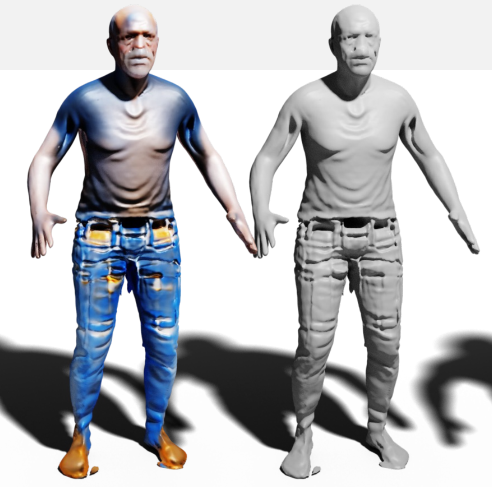
1. Superman
2. the face of Bill Gates



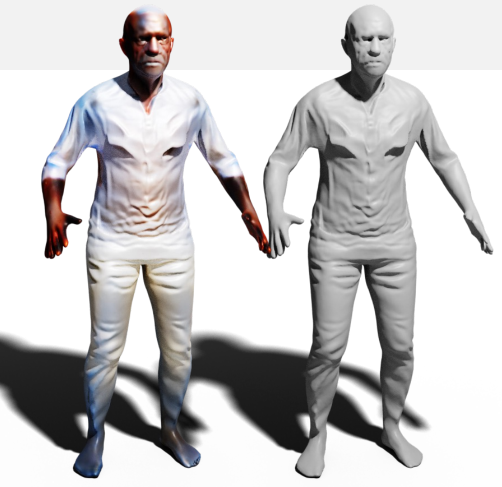
1. Iron Man
2. the face of Steve Jobs



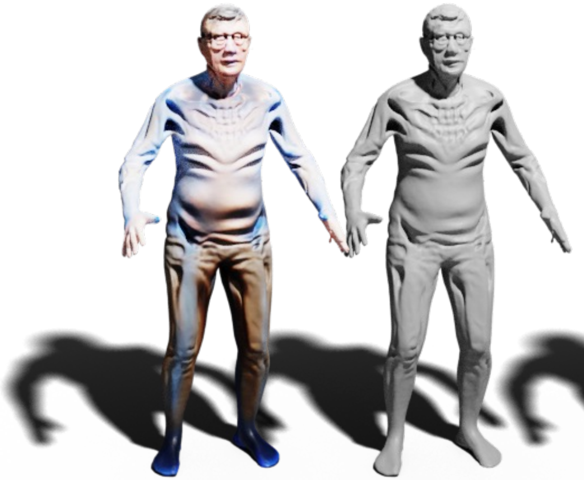
Steve Jobs in White Shirt



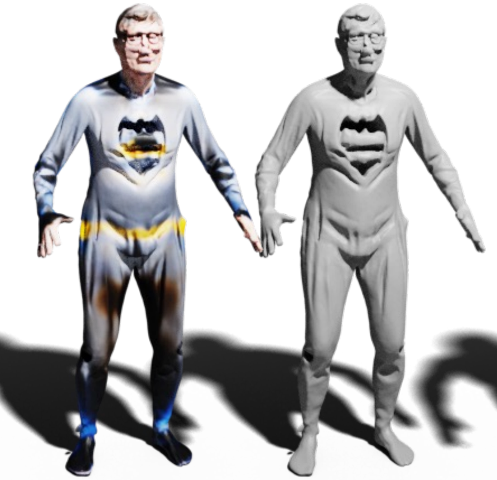
Man in Jeans



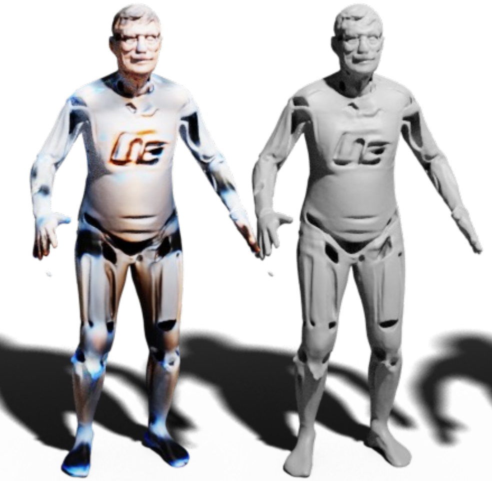
Man in White Shirt



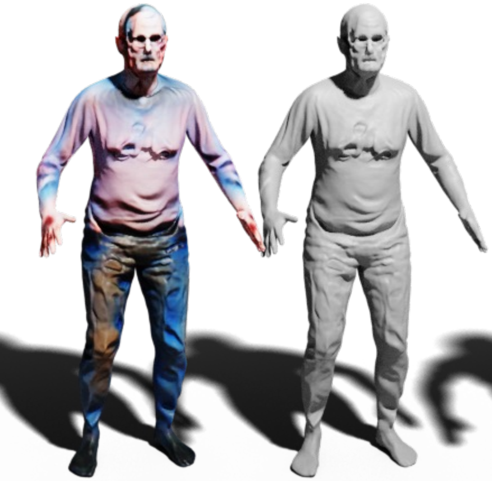
Alien Bill Gates



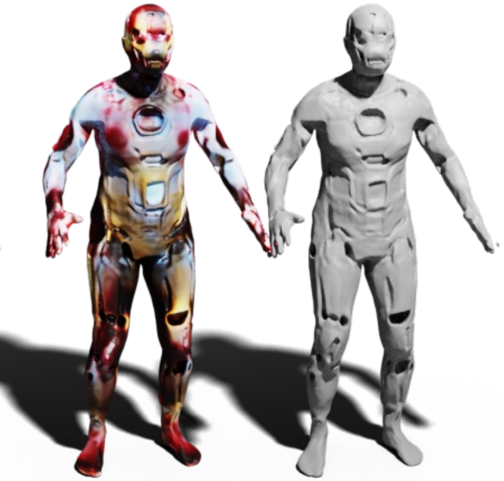
Bill Gates Wearing Batman Suit



Robot Bill Gates

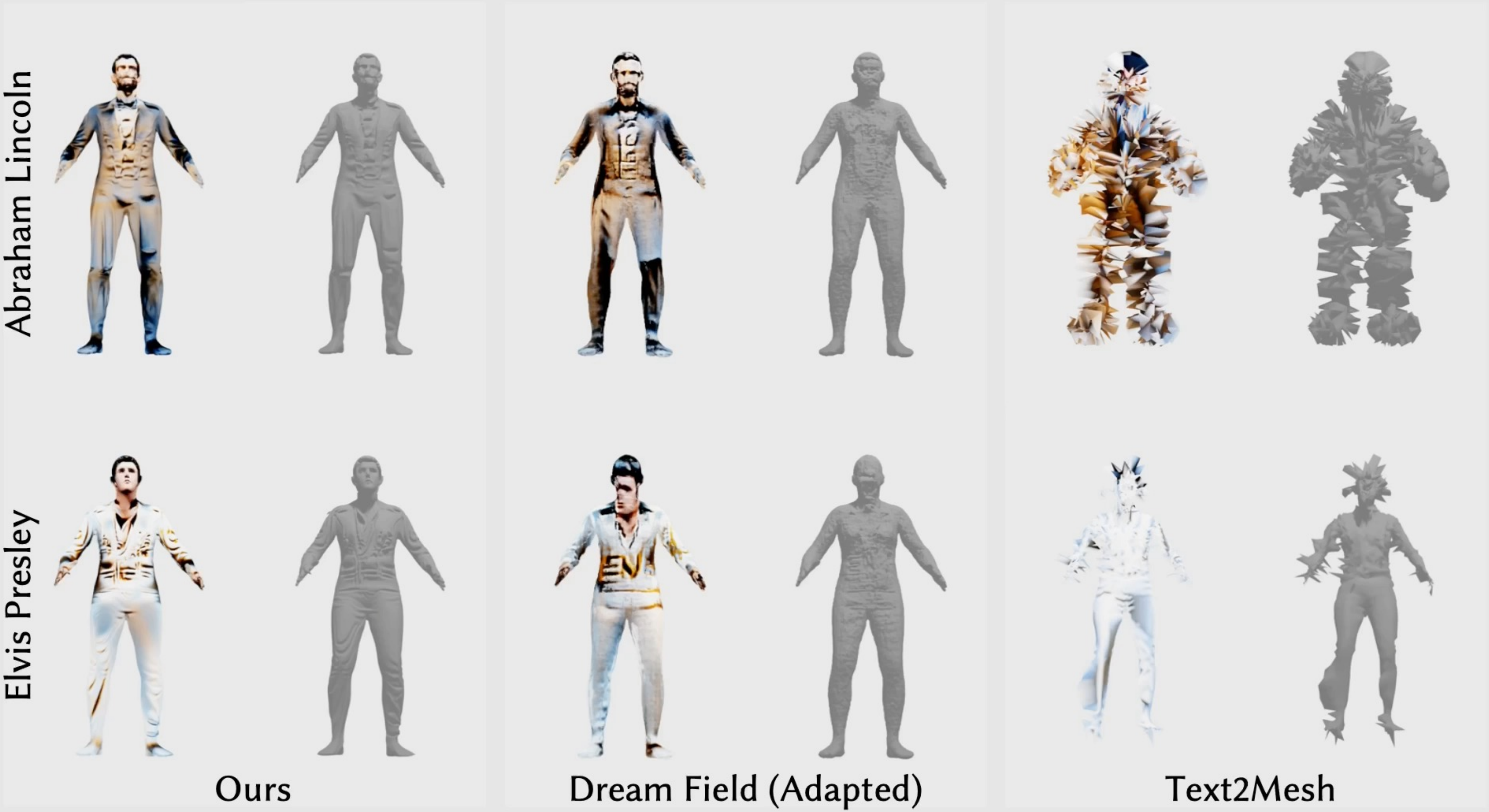


Zombie Steve Jobs

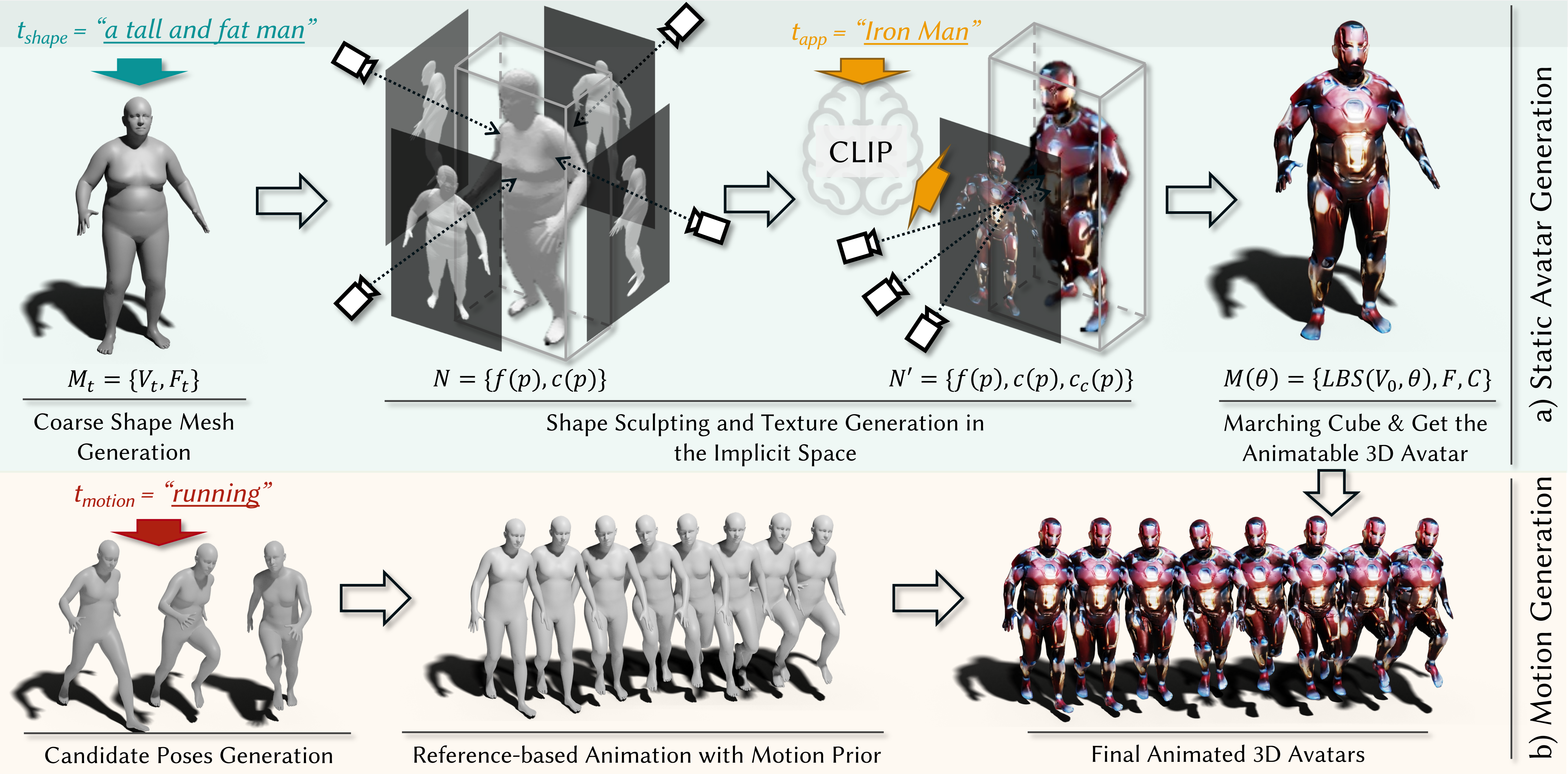


Zombie Iron Man

COMPARISON WITH BASELINE METHODS OF AVATAR GENERATION


Abraham Lincoln
Elvis Presley
Ours
Dream Field (Adapted)
Text2Mesh

AVATARCLIP: DETAILED PIPELINE



t_{shape} = *“a tall and fat man”*

$M_t = \{V_t, F_t\}$

Coarse Shape Mesh Generation

$N = \{f(p), c(p)\}$

CLIP

t_{app} = *“Iron Man”*

$N' = \{f(p), c(p), c_c(p)\}$

Shape Sculpting and Texture Generation in the Implicit Space

$M(\theta) = \{LBS(V_0, \theta), F, C\}$

Marching Cube & Get the Animatable 3D Avatar

a) Static Avatar Generation

t_{motion} = *“running”*

Candidate Poses Generation

Reference-based Animation with Motion Prior

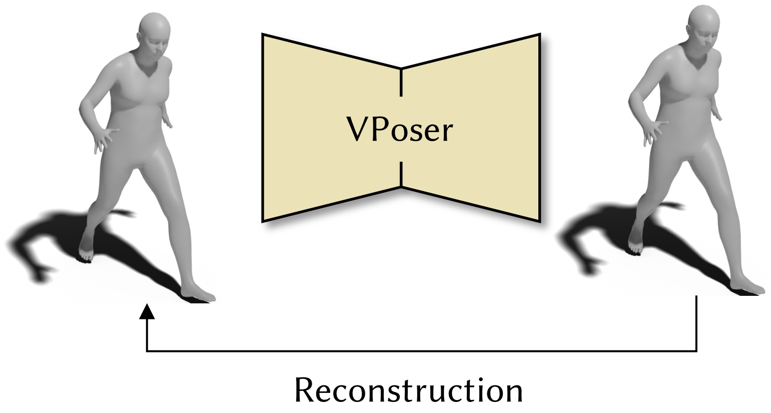
Final Animated 3D Avatars

b) Motion Generation

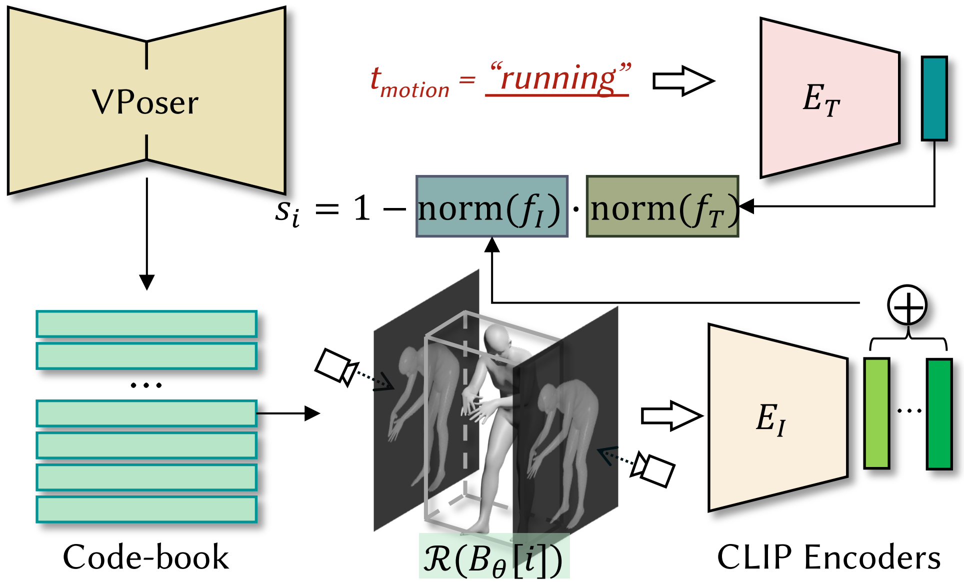
AVATARCLIP: CANDIDATE POSES GENERATION

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

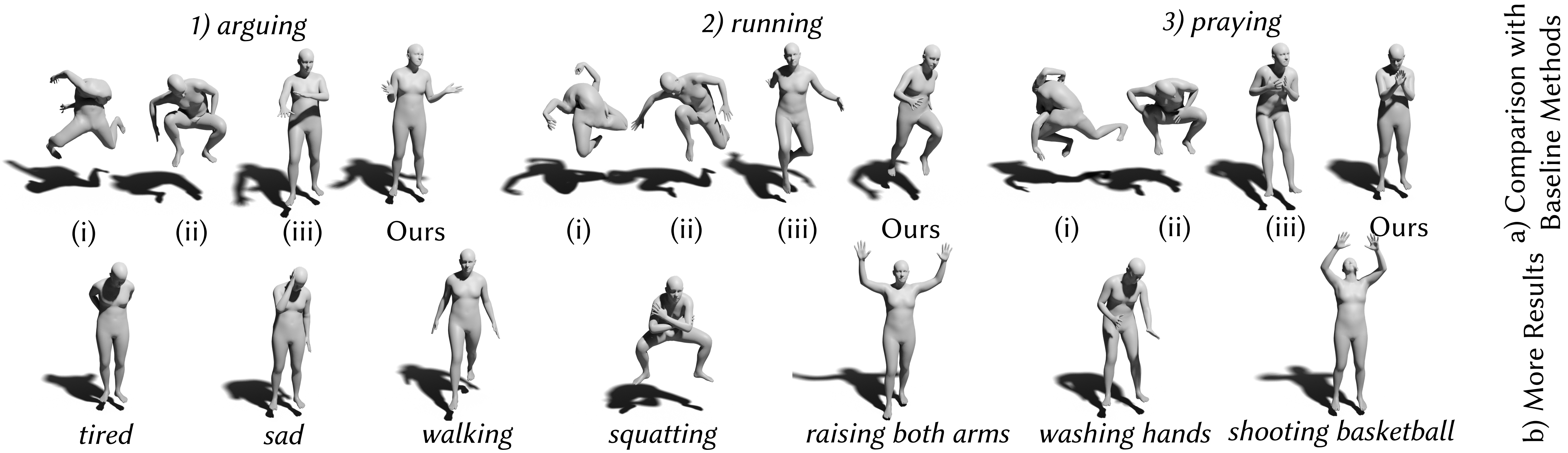
A) POSE VAE (VPOSER)



B) CLIP-GUIDED CANDIDATE POSES QUERY


VPoser
t_{motion} = "running"
E_T
$s_i = 1 - \text{norm}(f_I) \cdot \text{norm}(f_T)$
...
Code-book
$\mathcal{R}(B_\theta[i])$
E_I
...
CLIP Encoders

CANDIDATE POSE GENERATION RESULTS



(i) Direct optimization on SMPL parameter theta

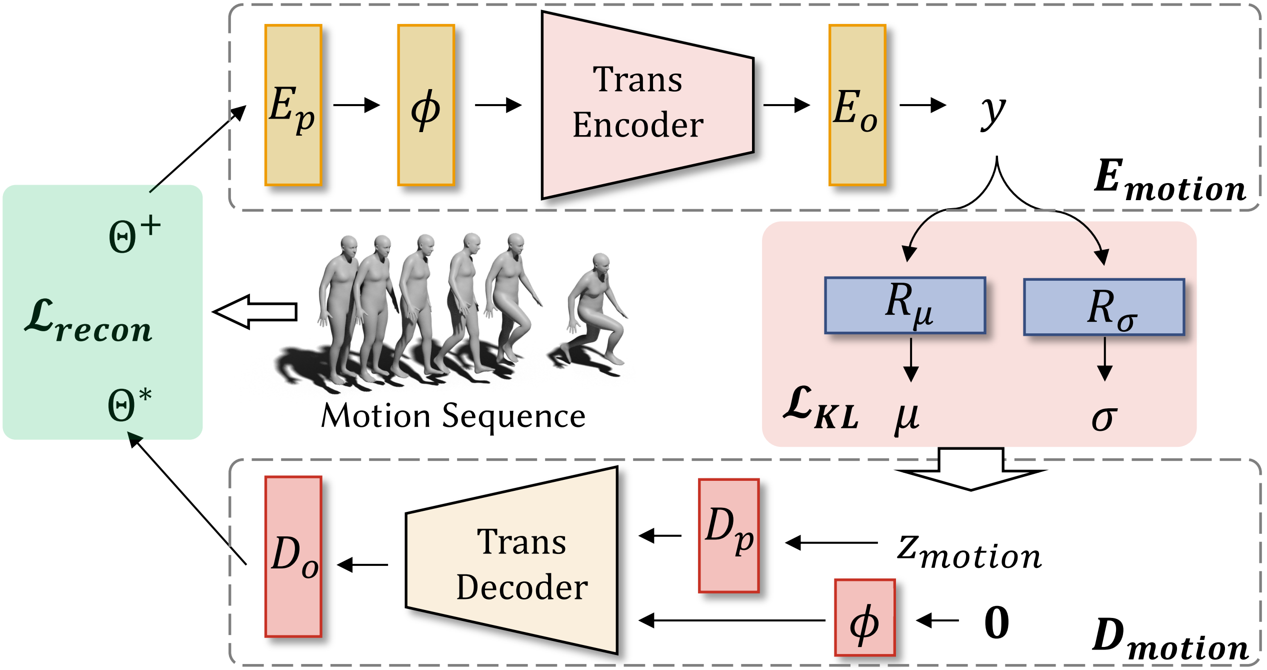
(ii) Direct optimization on VPoser latent code

(iii) Multi-Modal RealNVP

AVATARCLIP: REFERENCE-BASED ANIMATION

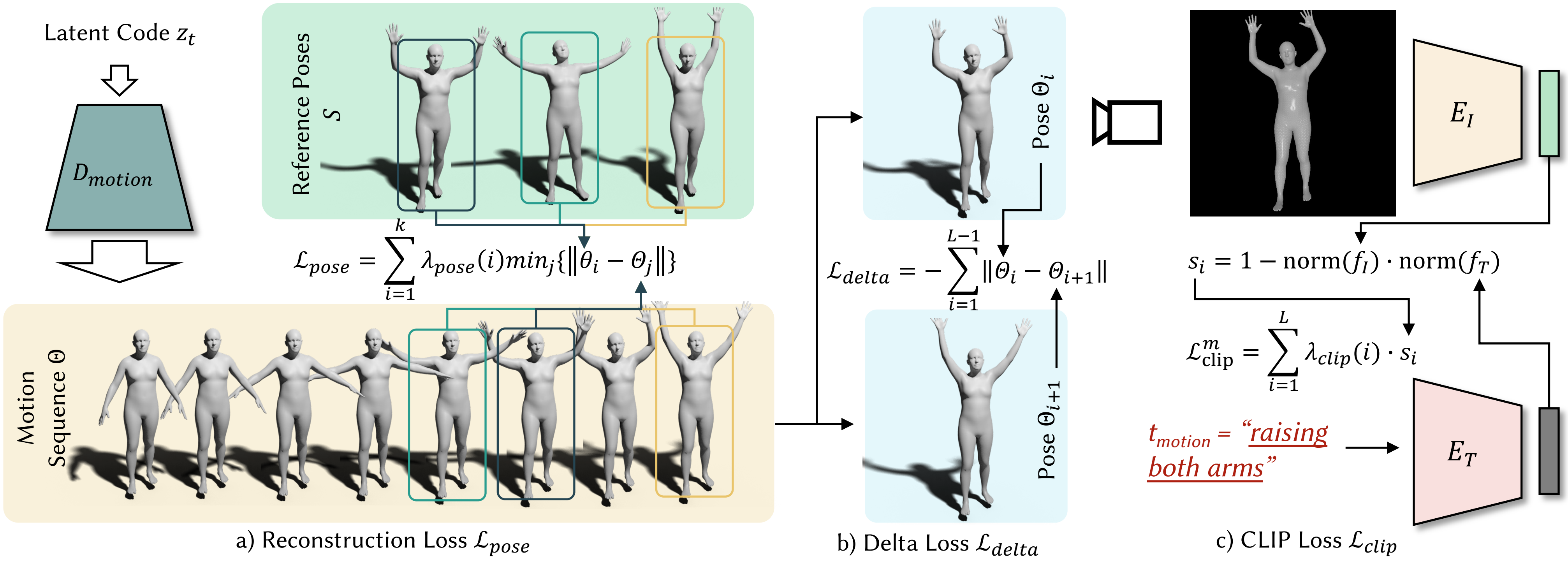
SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

A) MOTION VAE TRAINING

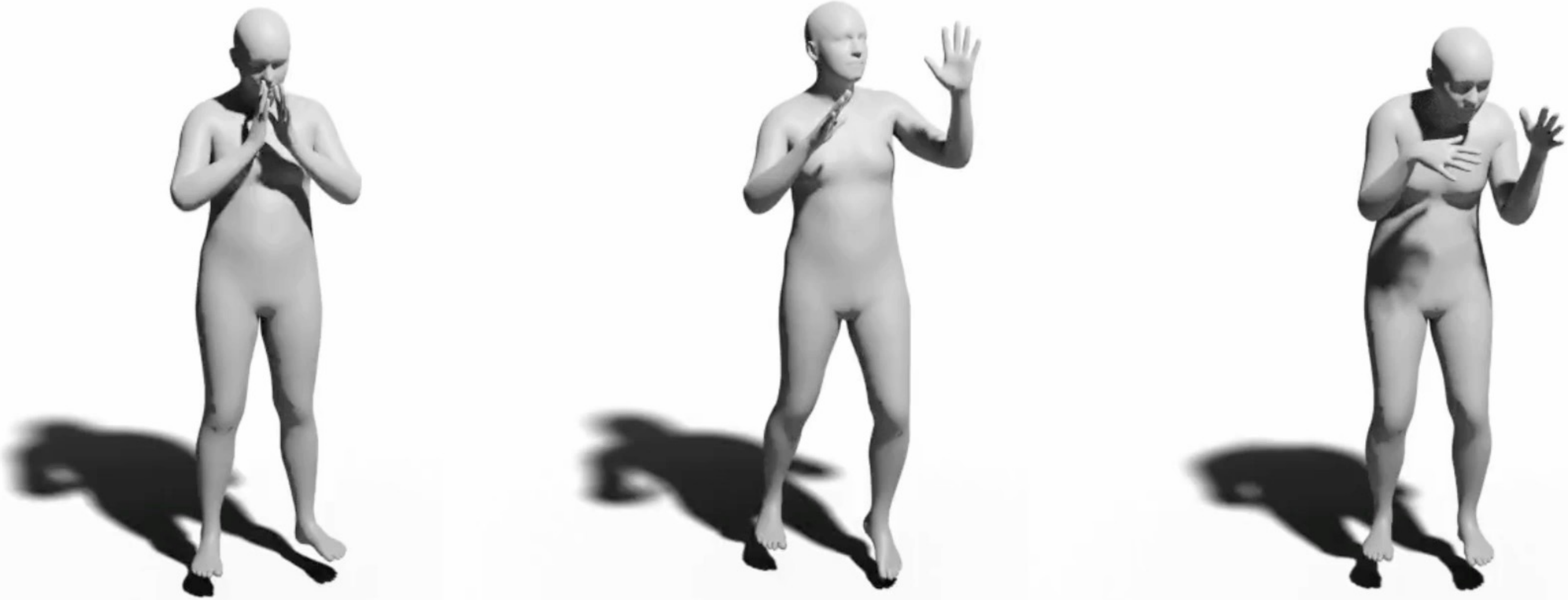


AVATARCLIP: REFERENCE-BASED ANIMATION (CONT.)

B) CLIP-GUIDED OPTIMIZATION ON THE MOTION VAE


Latent Code z_t
D_{motion}
Reference Poses S
$\mathcal{L}_{pose} = \sum_{i=1}^{k} \lambda_{pose}(i) min_j\{\|\theta_i - \Theta_j\|\}$
Motion Sequence Θ
a) Reconstruction Loss $\mathcal{L}_{pose}$
Pose Θ_i
$\mathcal{L}_{delta} = -\sum_{i=1}^{L-1} \|\Theta_i - \Theta_{i+1}\|$
Pose Θ_{i+1}
b) Delta Loss $\mathcal{L}_{delta}$
E_I
$s_i = 1 - \text{norm}(f_I) \cdot \text{norm}(f_T)$
$\mathcal{L}^m_{\text{clip}} = \sum_{i=1}^{L} \lambda_{clip}(i) \cdot s_i$
t_{motion} = "raising both arms"
E_T
c) CLIP Loss $\mathcal{L}_{clip}$

COMPARISONS OF MOTION GENERATION



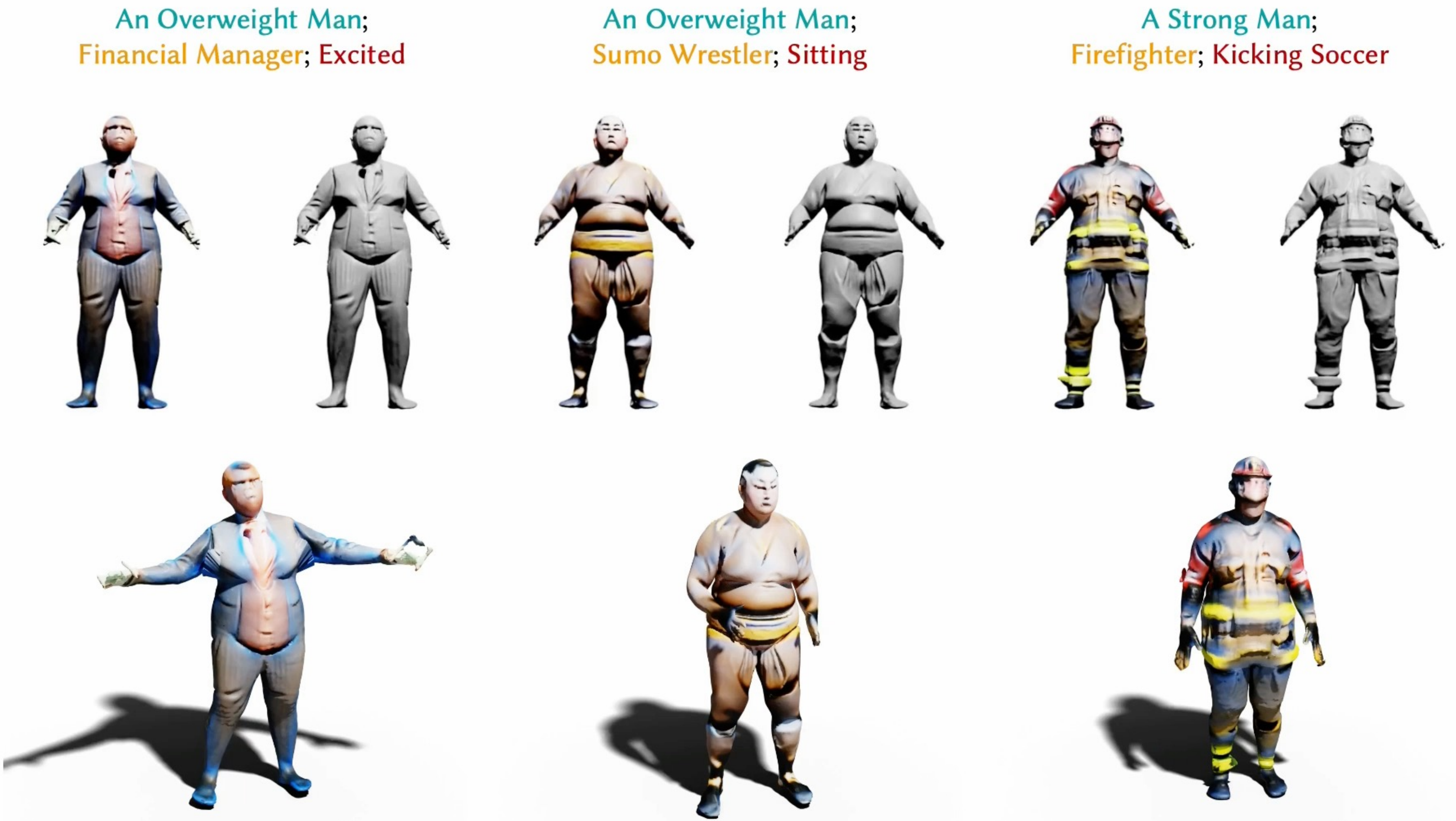
Direct Interpolation

Direct motion VAE optimization (Baseline)

Ours

Brushing Teeth

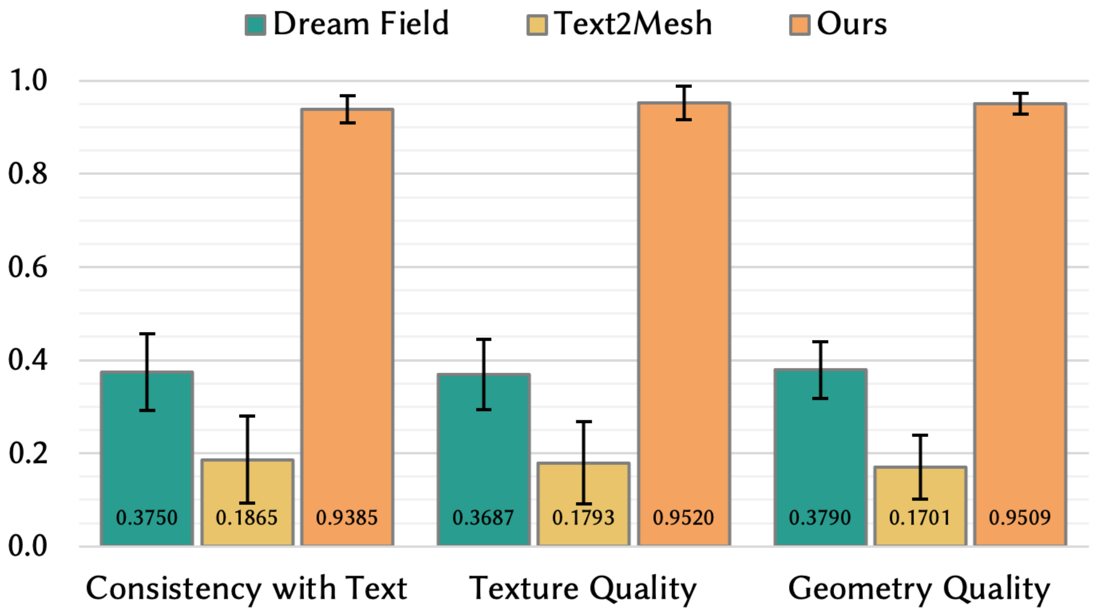
OVERALL RESULTS



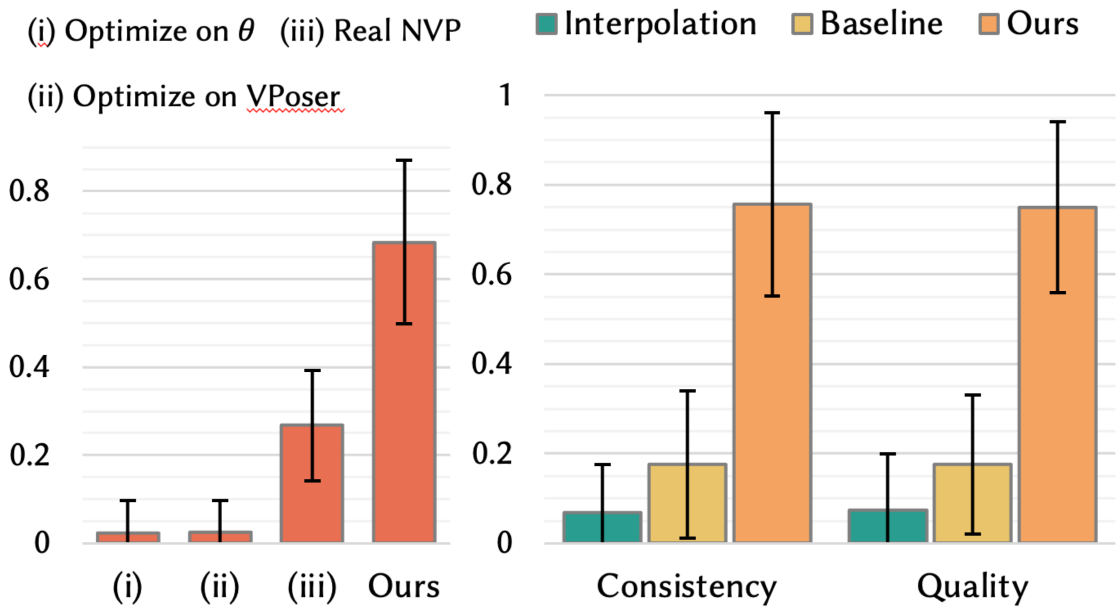
An Overweight Man; Financial Manager; Excited
An Overweight Man; Sumo Wrestler; Sitting
A Strong Man; Firefighter; Kicking Soccer

QUANTITATIVE RESULTS: USER STUDY

A) STATIC AVATAR GENERATION



B) MOTION GENERATION



(a) Candidate Poses Generation

(b) References-Based Animation

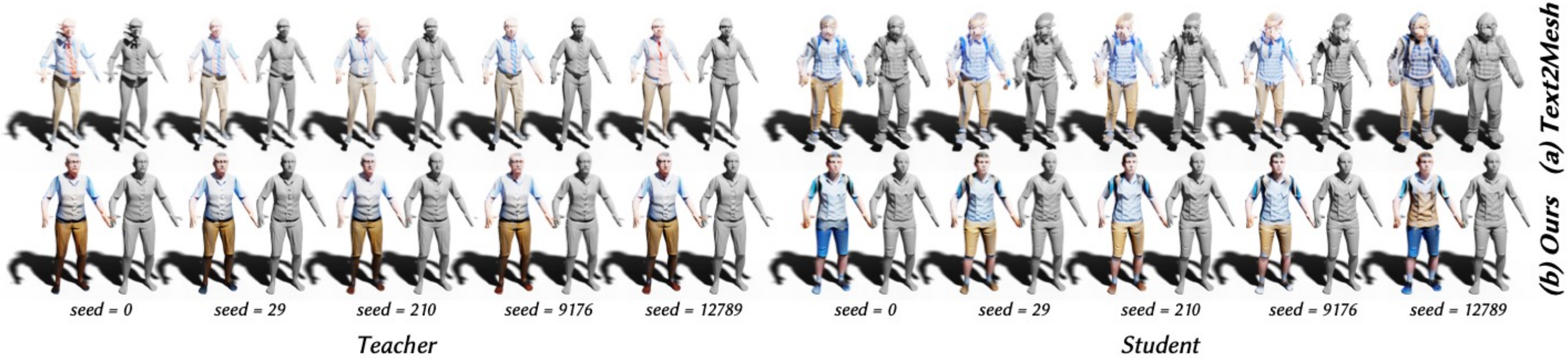
DISCUSSION

LIMITATIONS

- Low quality of generate avatar.
- Small variations across different runs.
- Hard to generate out-of-distribution poses.
- Difficult to generate stylized motions.

POTENTIAL NEGATIVE IMPACT

- Gender bias.
- Misused to make fake videos of celebrities.



THANK YOU

CODES ARE AVAILABLE

GitHub

Project Page