NEURAL INFORMATION PROCESSING SYSTEMS

Imperial College London

ID-to-3D

Expressive ID-guided 3D Heads via Score Distillation Sampling .

F. Babiloni, A.Lattas, J.Deng, S.Zafeiriou

Imperial College London

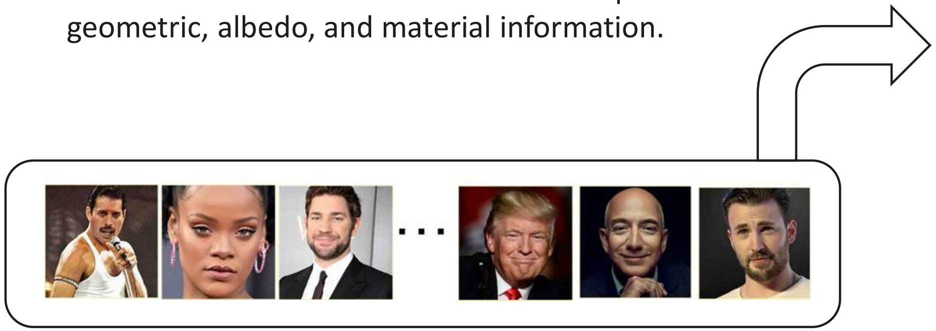
ID-to-3D: Expressive ID-guided 3D Heads via Score Distillation Sampling.

NEURAL INFORMATION PROCESSING SYSTEMS

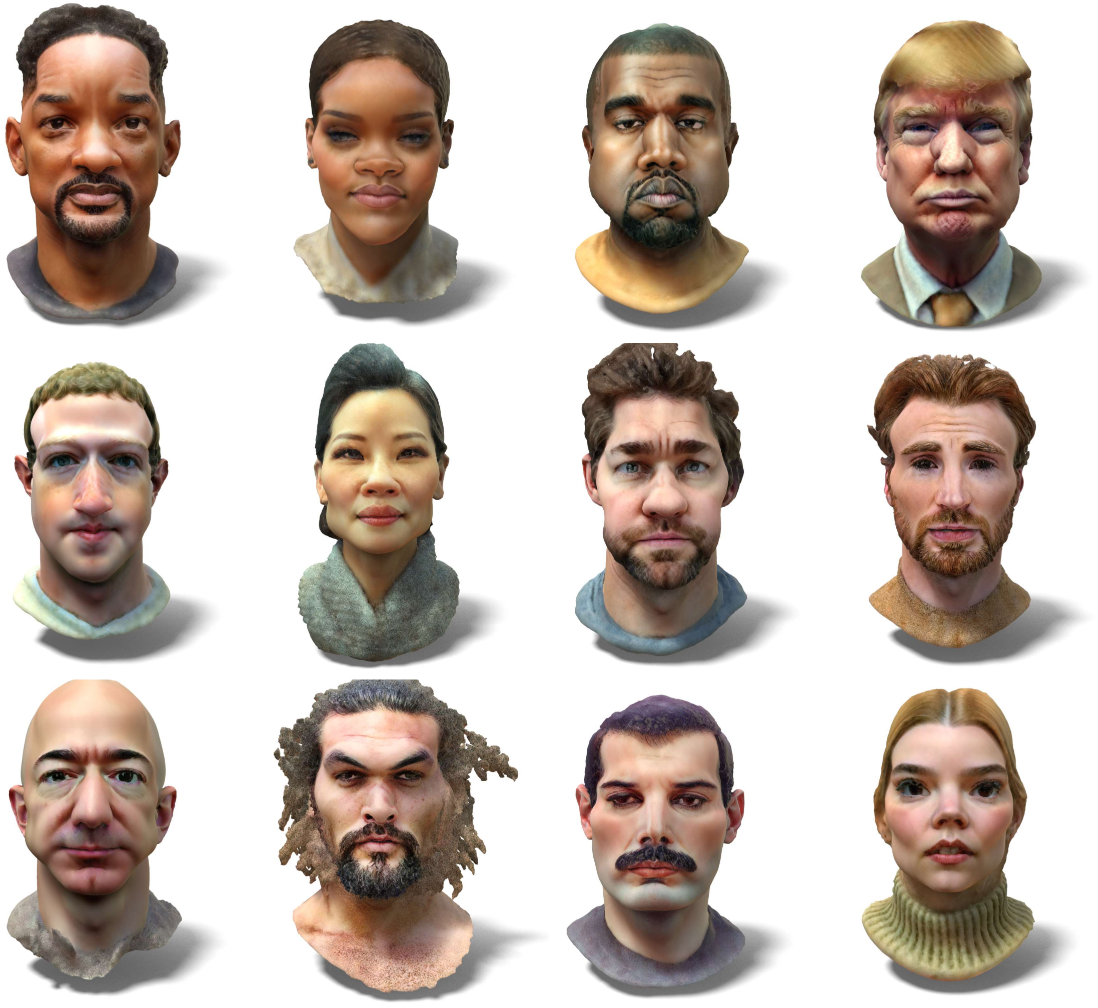
ArcFace-Conditioned 3D Head Asset Generation using SDS

- **Text-to-2D-Normals and Albedo Models**:
  We present a novel approach to creating ID-conditioned and expression-conditioned text-to-image models capable of generating realistically plausible normals and albedo images from a small set of 3D assets.

- **ID-Conditioned Expressive Heads**:
  We propose a neural parametric representation for the expressions that creates up to 13 unique and ID-consistent expressions captured by latent codes and associated with a set of 3D assets with separate geometric, albedo, and material information.



"In-the-wild" images



ID-consistent 3D assets

3D Head Optimization Objective

Score Distillation Sampling Pipeline Optimization Objective

$$\min_{\theta} D_{KL}(\underbrace{q^{\theta}(\mathbf{x}_0|y_{\text{text}})}_{\text{distribution of renderings}} \,\|\, \underbrace{p(\mathbf{x}_0|y_{\text{text}})}_{\text{target distribution}}).$$

- ⊗ Target distribution drift.
- ⊗ General guidance model does not separate texture/geometry
- ⊗ Using textual prompt lacks granularity.

3D Head Optimization Objective

ID to 3D Optimization Objective

$$\min_{\theta_g,\theta_a} \underbrace{D_{KL}(q^{\theta_g}(\mathbf{z}_0^n|\mathbf{c}, y_{\text{text}}, y_{\text{exp}}, y_{\text{id}}) \parallel p(\mathbf{z}_0^n|\mathbf{c}, y_{\text{text}}, y_{\text{exp}}, y_{\text{id}}))}_{geometry\ generation\ objective}$$

$$+ \underbrace{D_{KL}(q^{\theta_a}(\mathbf{z}_0^a|\mathbf{c}, \mathbf{l}, y_{\text{text}}, y_{\text{exp}}, y_{\text{id}}) \parallel p(\mathbf{z}_0^a|\mathbf{c}, y_{\text{text}}, y_{\text{exp}}, y_{\text{id}}))}_{texture\ generation\ objective}.$$

θ_g parameterization of the 3D geometry.

θ_a parameterization of the 3D textures

$\mathbf{z}_0^a$ albedo textures

$\mathbf{z}_0^n$ normal maps

$\mathbf{l}$ $\mathbf{c}$ lighting camera

y_{text} textual

y_{id} identity

y_{exp} expression

2D Guidance

- Geometry / Texture adapted self-attention:

$$\mathbf{Z}^n_{\mathrm{SA}} = \mathrm{Att}(\mathbf{Q}^n, \mathbf{K}^n, \mathbf{V}^n). \qquad \mathbf{Q}^n = \mathbf{X}\mathbf{W}_Q + \mathbf{X}\mathbf{W}^n_Q, \mathbf{K}^n = \mathbf{X}\mathbf{W}_K + \mathbf{X}\mathbf{W}^n_K, \mathbf{V}^n = \mathbf{X}\mathbf{W}_V + \mathbf{X}\mathbf{W}^n_V$$

- Identity and expression adapted cross-attention

$$\mathbf{Z}^n_{\mathrm{CA}} = \mathrm{Att}(\mathbf{Q}, \mathbf{K}^{\mathrm{text}}, \mathbf{V}^{\mathrm{text}}) + \lambda_{id} \cdot \mathrm{Att}(\mathbf{Q}, \mathbf{K}^{\mathrm{id}}, \mathbf{V}^{\mathrm{id}}) + \lambda_{exp} \cdot \mathrm{Att}(\mathbf{Q}, \mathbf{K}^{\mathrm{exp}}, \mathbf{V}^{\mathrm{exp}})$$

- Efficient Finetuning

$$L_{\mathrm{simple}} = \mathbb{E}_{\mathbf{z}^n_0, \epsilon, \boldsymbol{c}, t, \mathbf{y}_{\mathrm{text}}, \mathbf{y}_{\mathrm{id}}, \mathbf{y}_{\mathrm{exp}}} \| \epsilon - \epsilon_{\phi_g}(\mathbf{z}^n_t, t, \boldsymbol{c}, \mathbf{y}_{\mathrm{text}}, \mathbf{y}_{\mathrm{id}}, \mathbf{y}_{\mathrm{exp}}) \|^2$$

ArcFace

CLIP exp textual descriptors

ϕ_g	2D prior for geometry
ϕ_a	2D prior fortextures
$\mathbf{y}_{\mathrm{id}}$	Arcface identity embeddings
$\mathbf{y}_{\mathrm{exp}}$	CLIP embeddings textual descriptor of FaceWareHouse
$\mathbf{c}$	textual descriptor of the camera poses

3D Generation

- Geometry Generation

$$\nabla \mathcal{L}_{SDS}(\theta_g) = \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\omega(t)(\epsilon_{\phi_g}(\mathbf{z}_t^n, \mathbf{y}_{\text{id}}, \mathbf{y}_{\text{exp}}, \mathbf{y}_{\text{text}}, t) - \epsilon)\frac{\partial g(\theta_g, \mathbf{c})}{\partial \theta_g}\right]$$

$$\theta_g = [\mathbf{k}^n_{\text{exp}}, \psi_g]$$

θ_g geometry model

$\Psi_g(\Gamma, \mathbf{k}^n_{\text{exp}})$ Transformer DMTET

$\mathbf{k}^n_{\text{exp}} \in \mathbb{R}^{d_{exp}}$ latent code

Γ deformable tetrahedral grid

$g(\theta_g, \mathbf{c})$ rendered normals

ϕ_g 2D prior for geometry

- Appearance Generation

$$\nabla \mathcal{L}_{SDS}(\theta_a) = \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\omega(t)(\epsilon_{\phi_a}(\mathbf{z}_t^a, \mathbf{y}_{\text{id}}, \mathbf{y}_{\text{exp}}, \mathbf{y}_{\text{text}}, t) - \epsilon)\frac{\partial g(\theta_a, \mathbf{c}, \mathbf{l})}{\partial \theta_a}\right]$$

$$\theta_a = [\mathbf{k}^a_{\text{exp}}, \psi_a]$$

θ_a texture appearance

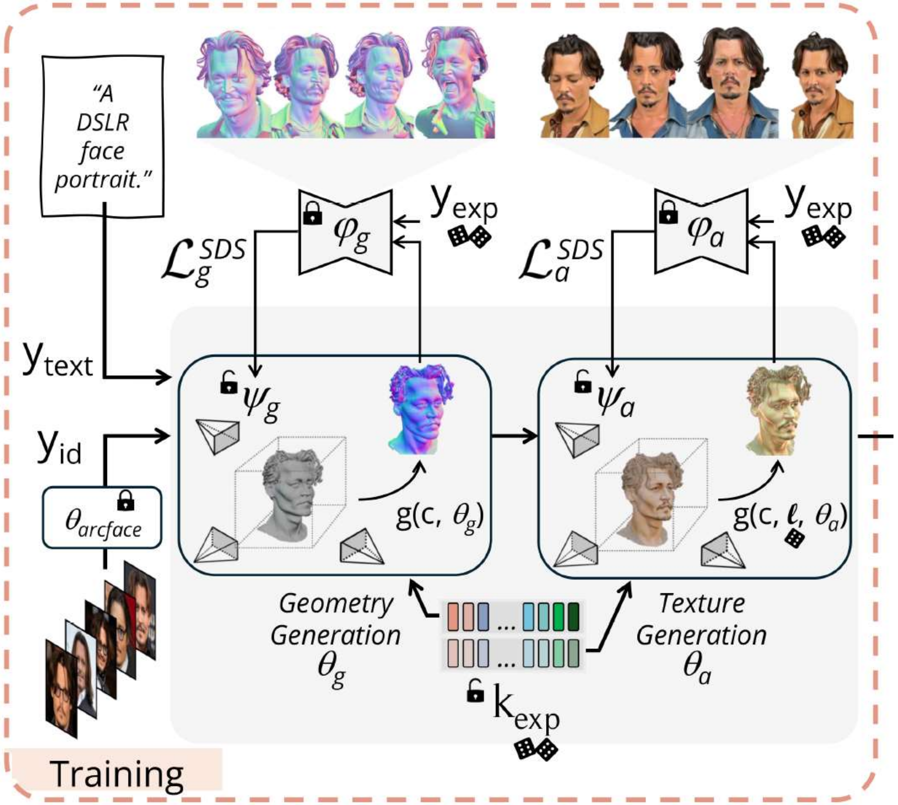
$\Psi_a(\mathbf{k}^a_{\text{exp}})$ Transformer

roughness
albedo
specularity

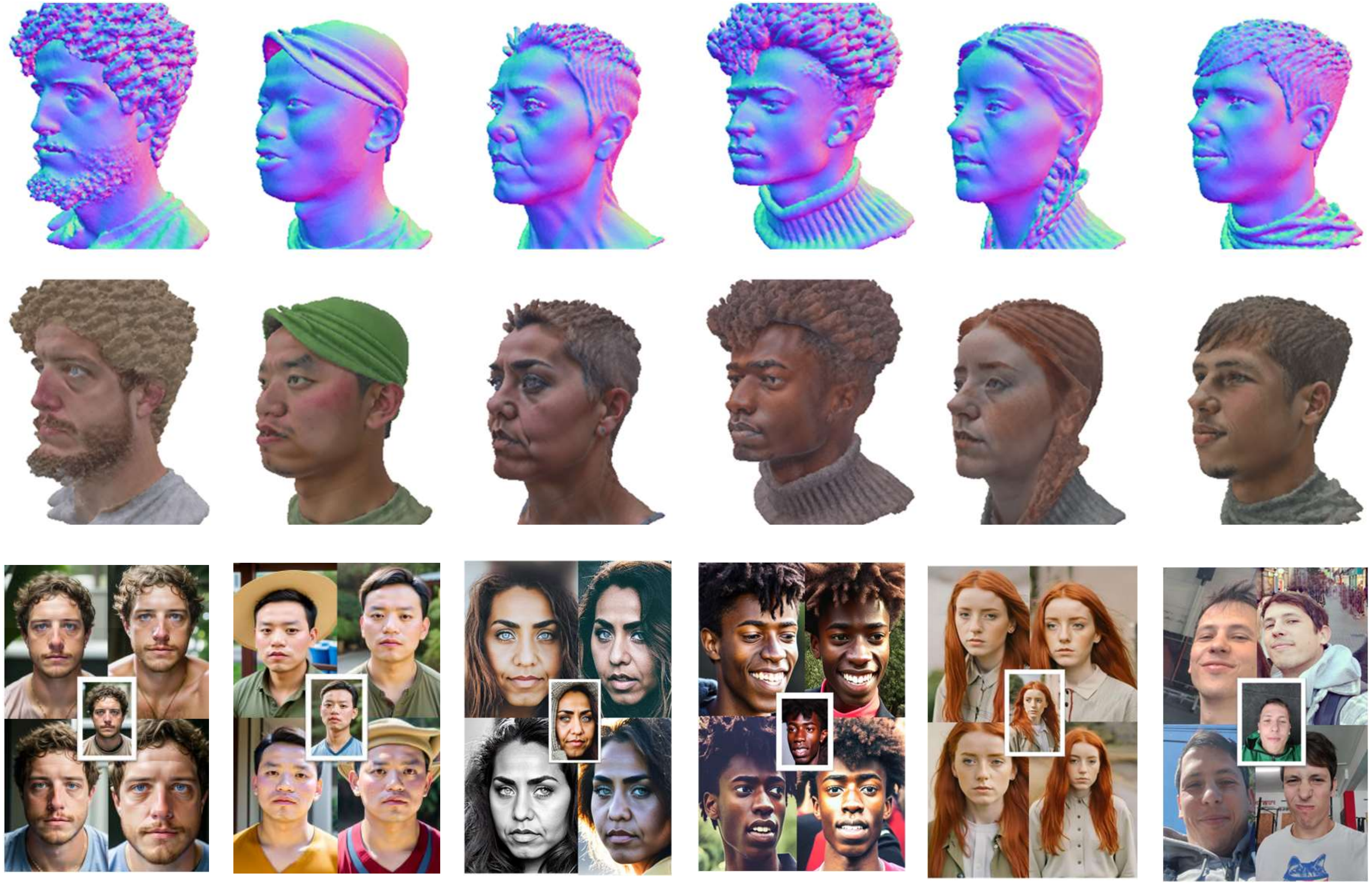
$\mathbf{k}^a_{\text{exp}} \in R^{d_{exp}}$ latent code

$g(\theta_a, \mathbf{c}, \mathbf{l})$ rendered pseudo-albedo

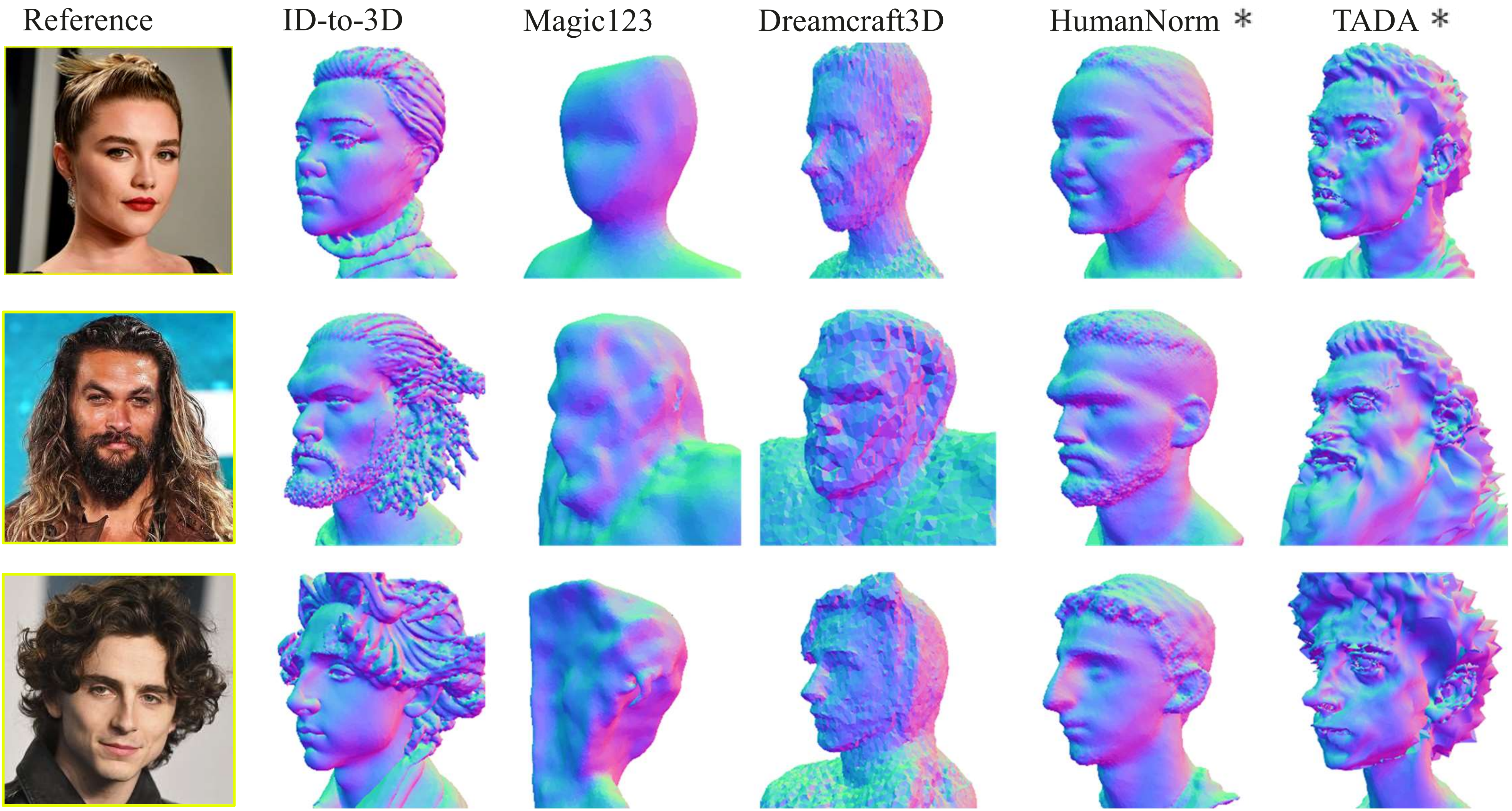
ϕ_a 2D prior fortextures



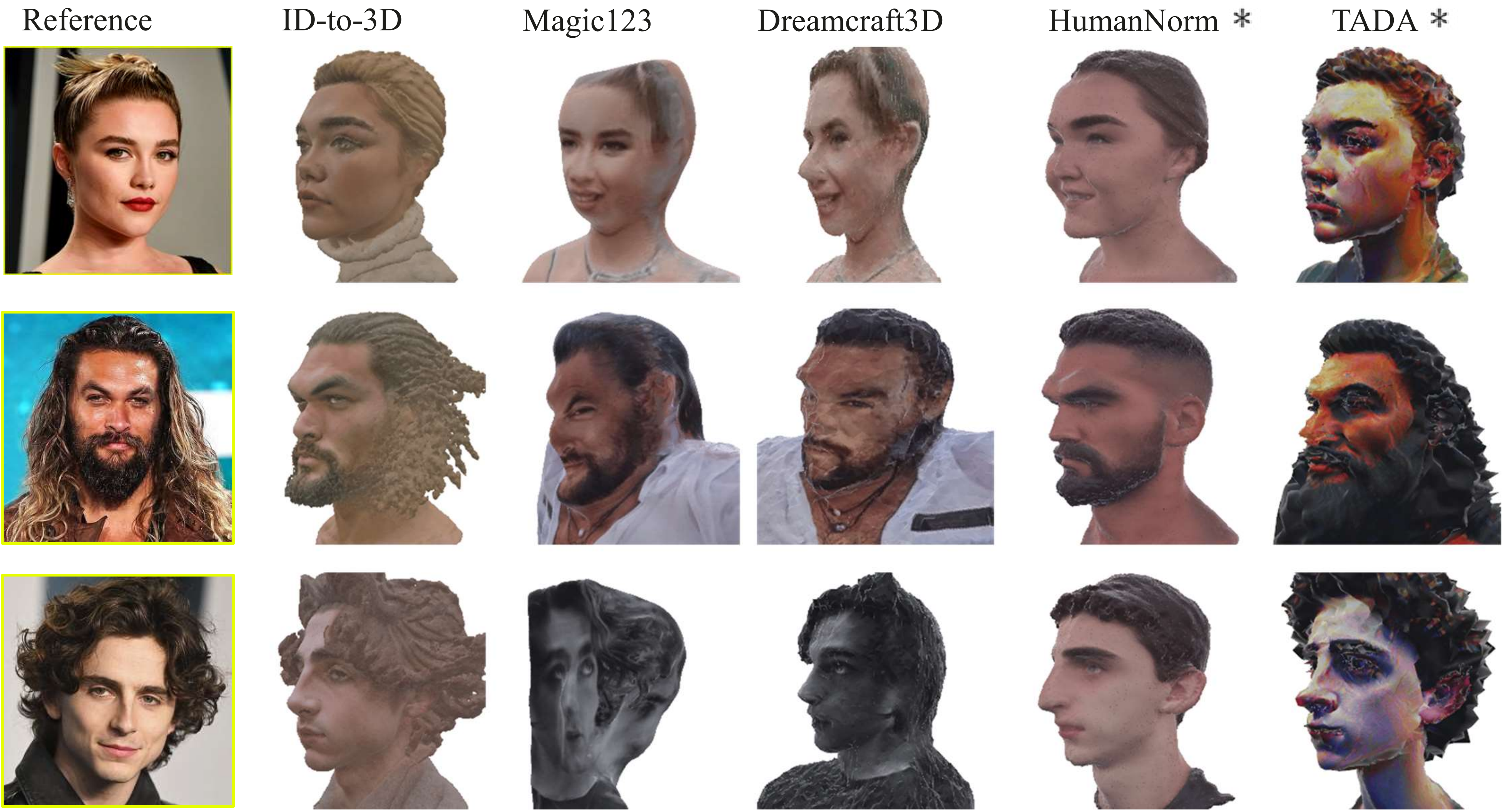
Results: ID-consistent texture and geometry generation



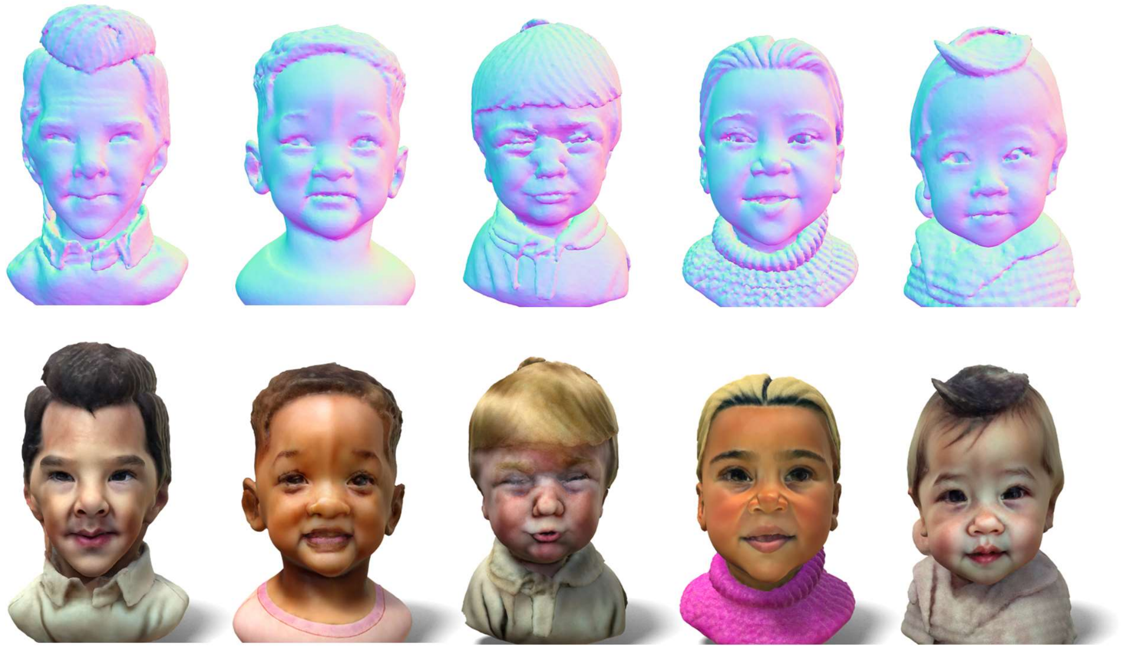
Results: Comparison with SotA SDS methods



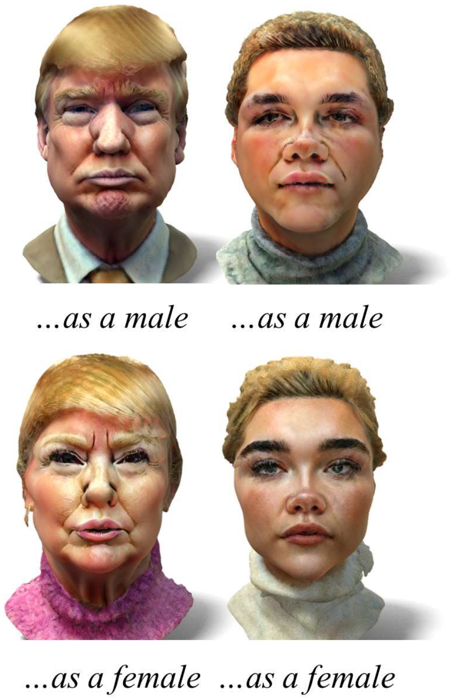
Results: Comparison with SotA SDS methods



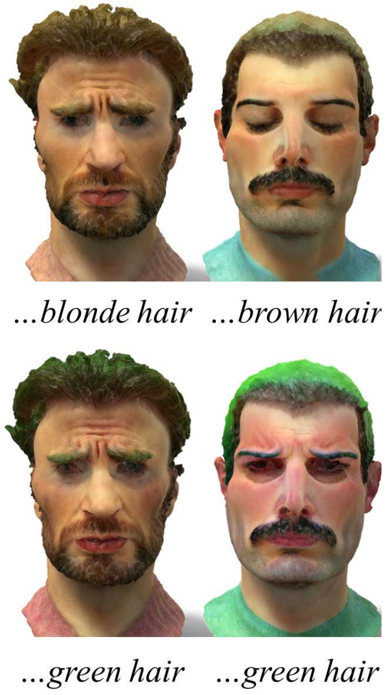
Results: ID-consistent text-based customization



"...as a cute baby"

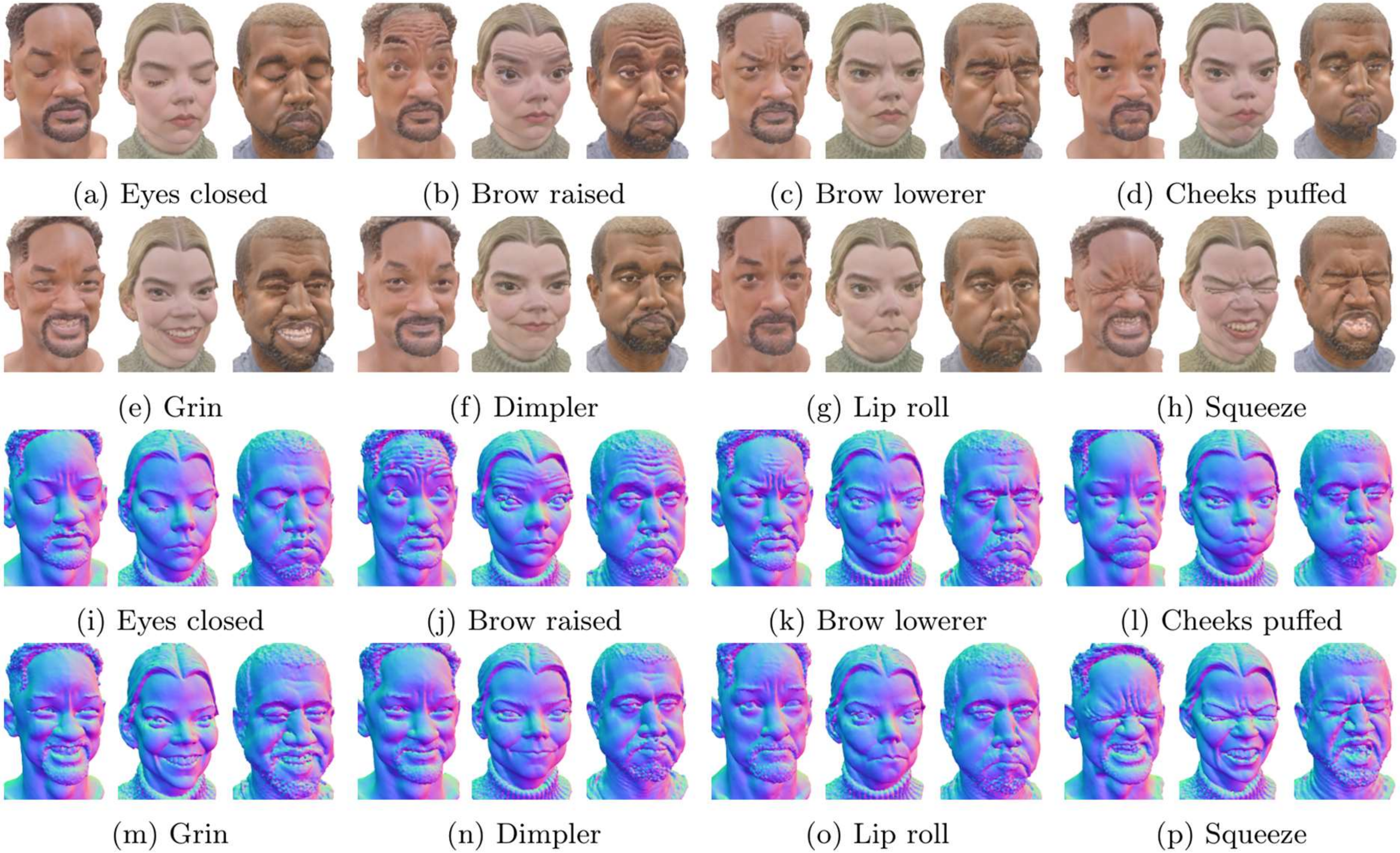


Geometry Editing



Texture Editing

Results: Expressive ID-conditioned Generation



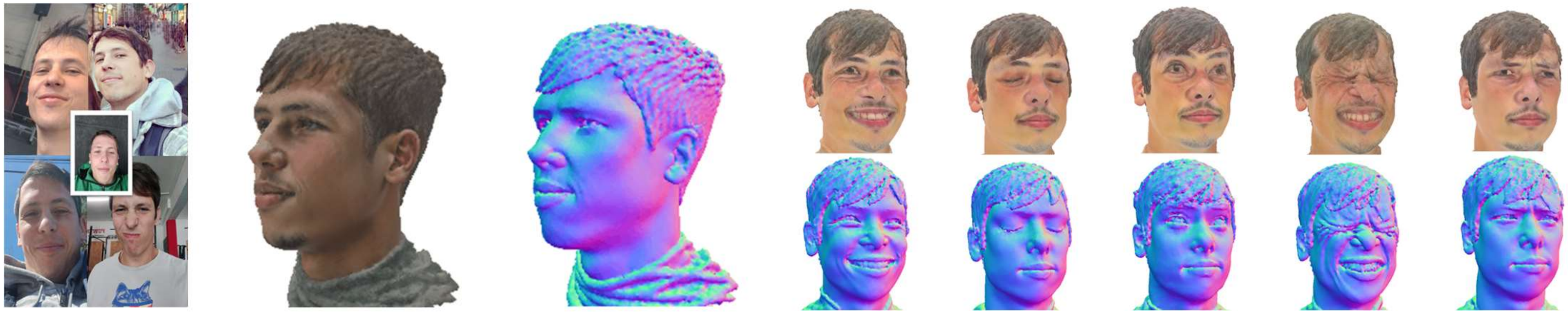
(a) Eyes closed (b) Brow raised (c) Brow lowerer (d) Cheeks puffed

(e) Grin (f) Dimpler (g) Lip roll (h) Squeeze

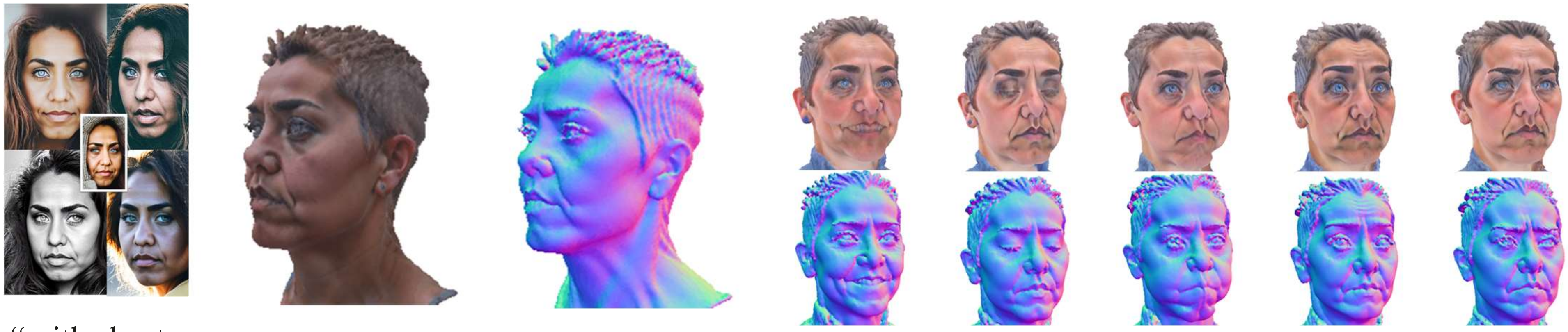
(i) Eyes closed (j) Brow raised (k) Brow lowerer (l) Cheeks puffed

(m) Grin (n) Dimpler (o) Lip roll (p) Squeeze

Results: Expressive ID-conditioned Generation

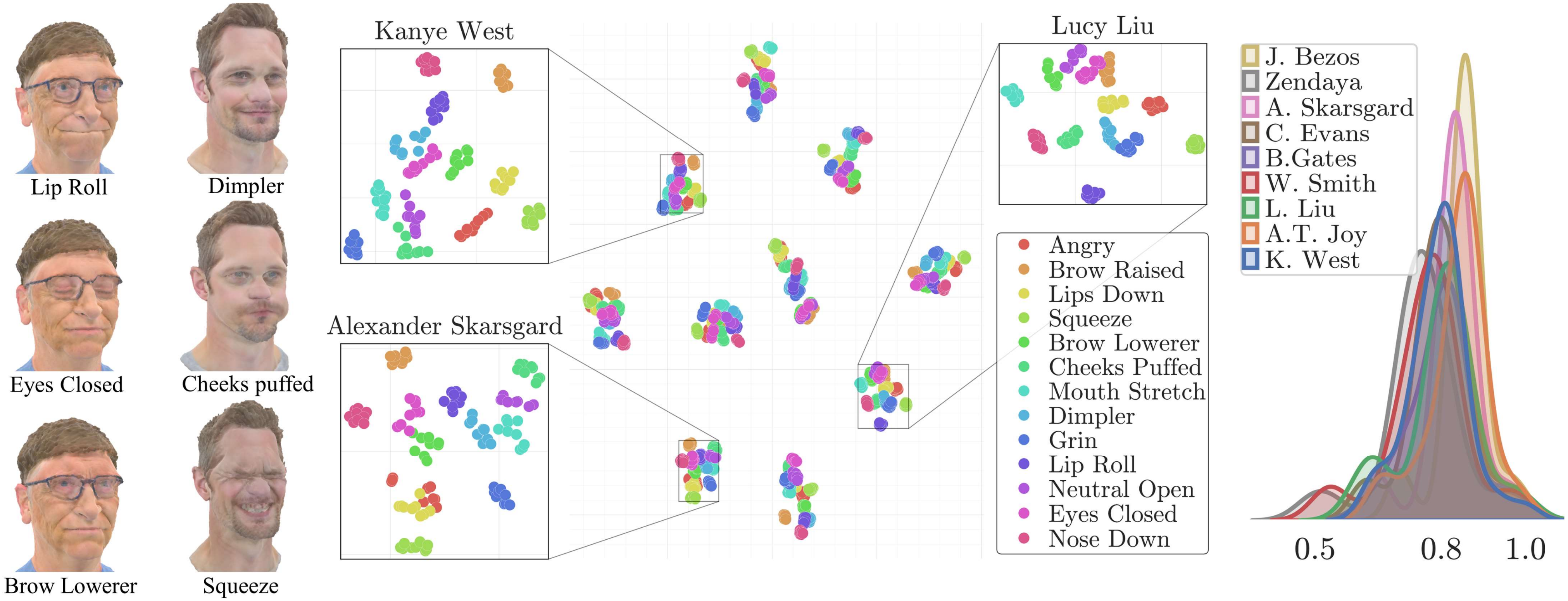


-

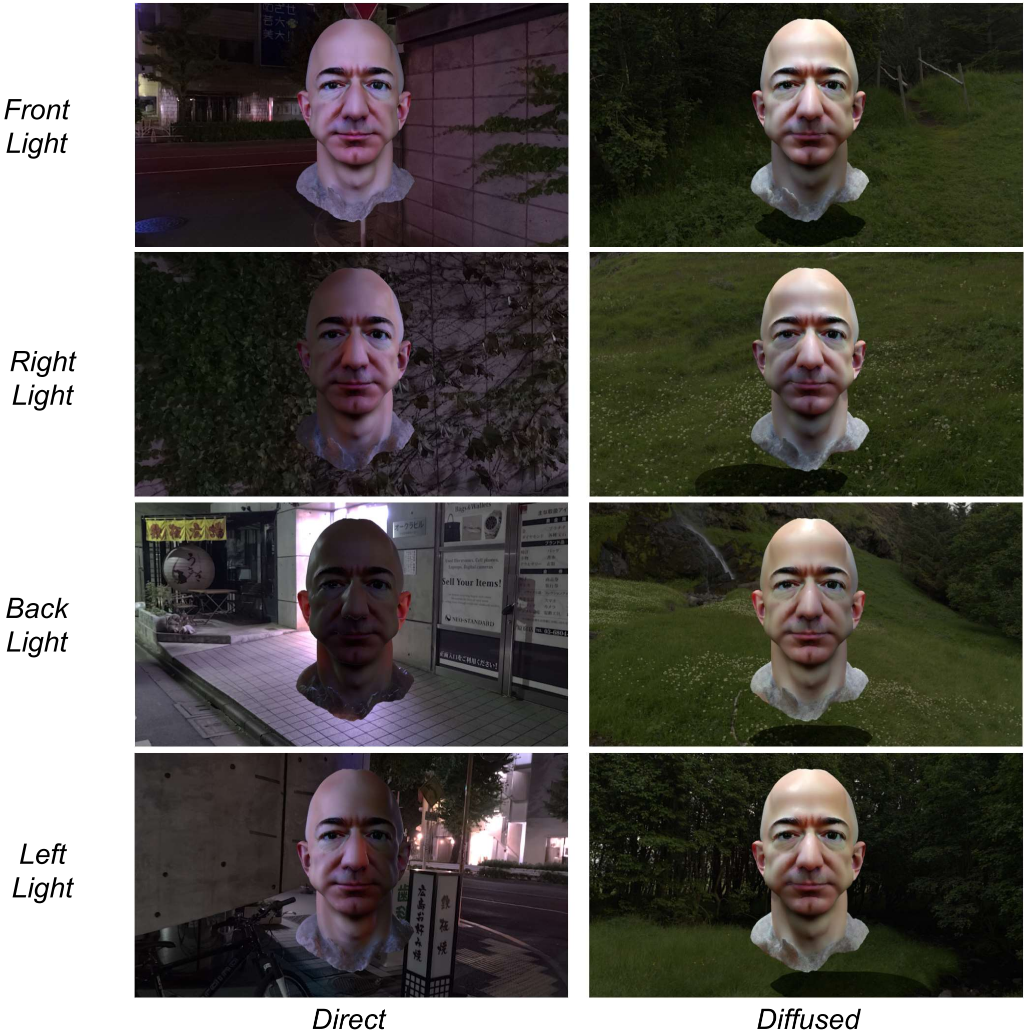


“with short buzzcut hairstyle”

Results: Expressive ID-conditioned Generation



Results: Relighting



THANKS!