Microsoft

Flow-based 3D Avatar Generation from Sparse Observations

Sadegh Aliakbarian
saliakbarian@microsoft.com

In collaboration with: Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Tom Cashman

Goal

Given the signal from head-mounted devices (HMDs), the goal is to generate realistic and faithful full-body avatar poses to represent people in mixed reality scenarios.

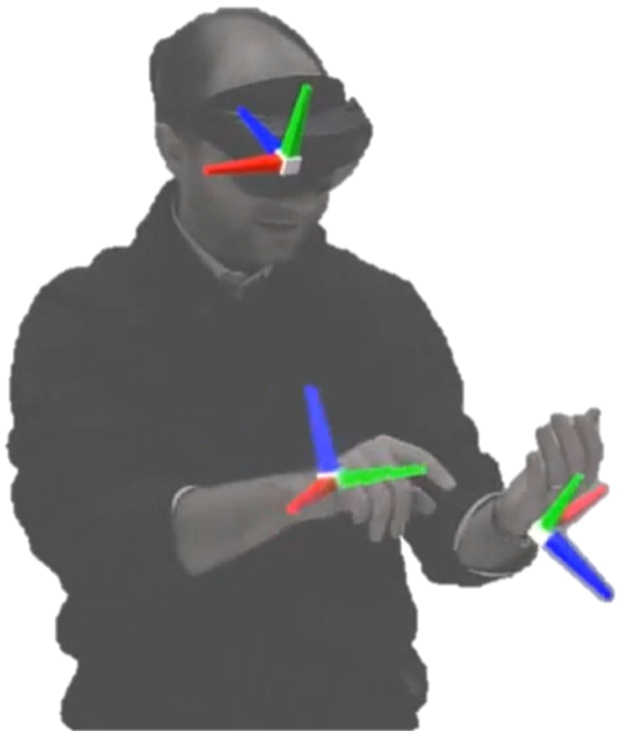
Overview

A user is wearing a head-mounted device (HMD), e.g., HoloLens2

HMD provides the location and orientation of head and hands

Our approach predicts full-body avatar pose given HMD signal

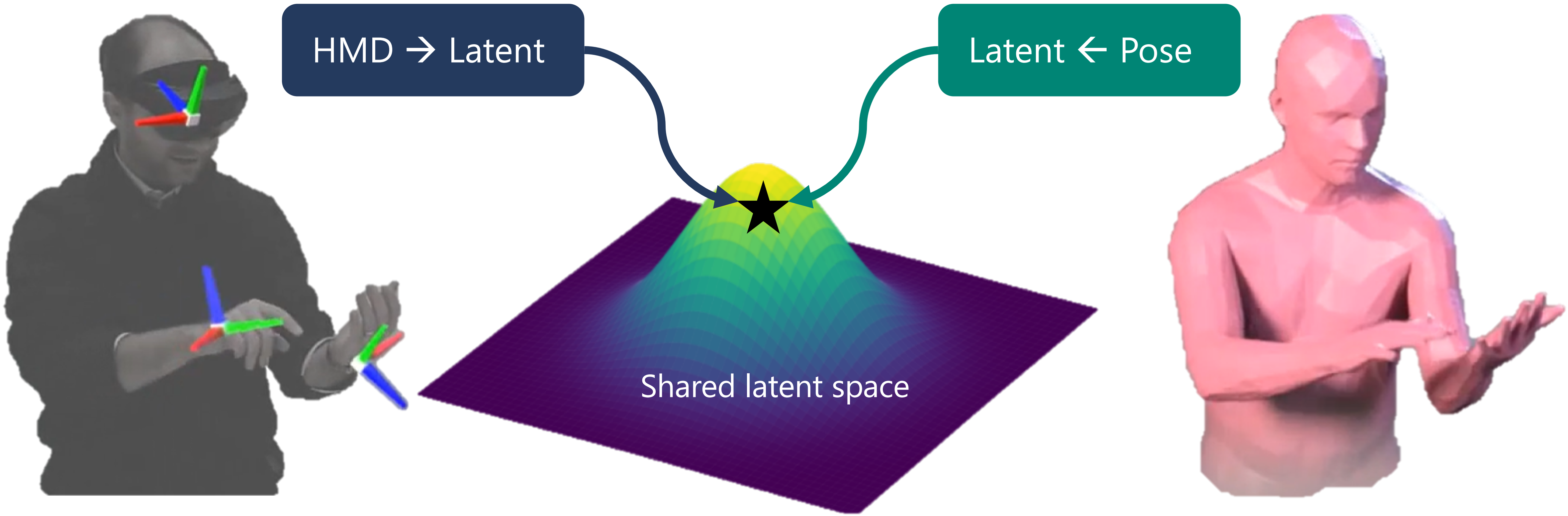
Idea



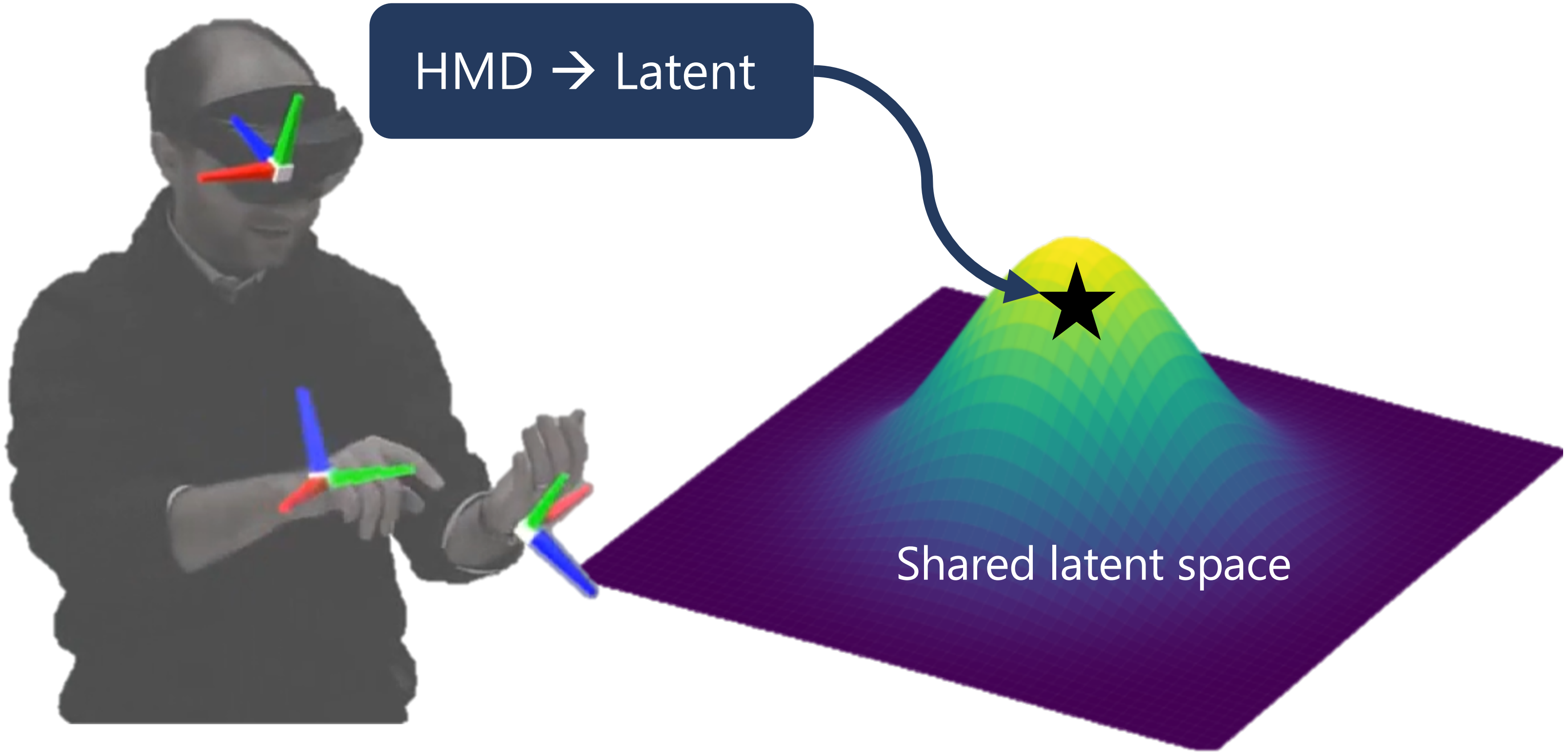
HMD signal and 3D pose
represent same person

Can we learn a common representation?

Idea



Idea



HMD → Latent properties

Deal with sparse observation

Deal with uncertainty

Expressive enough

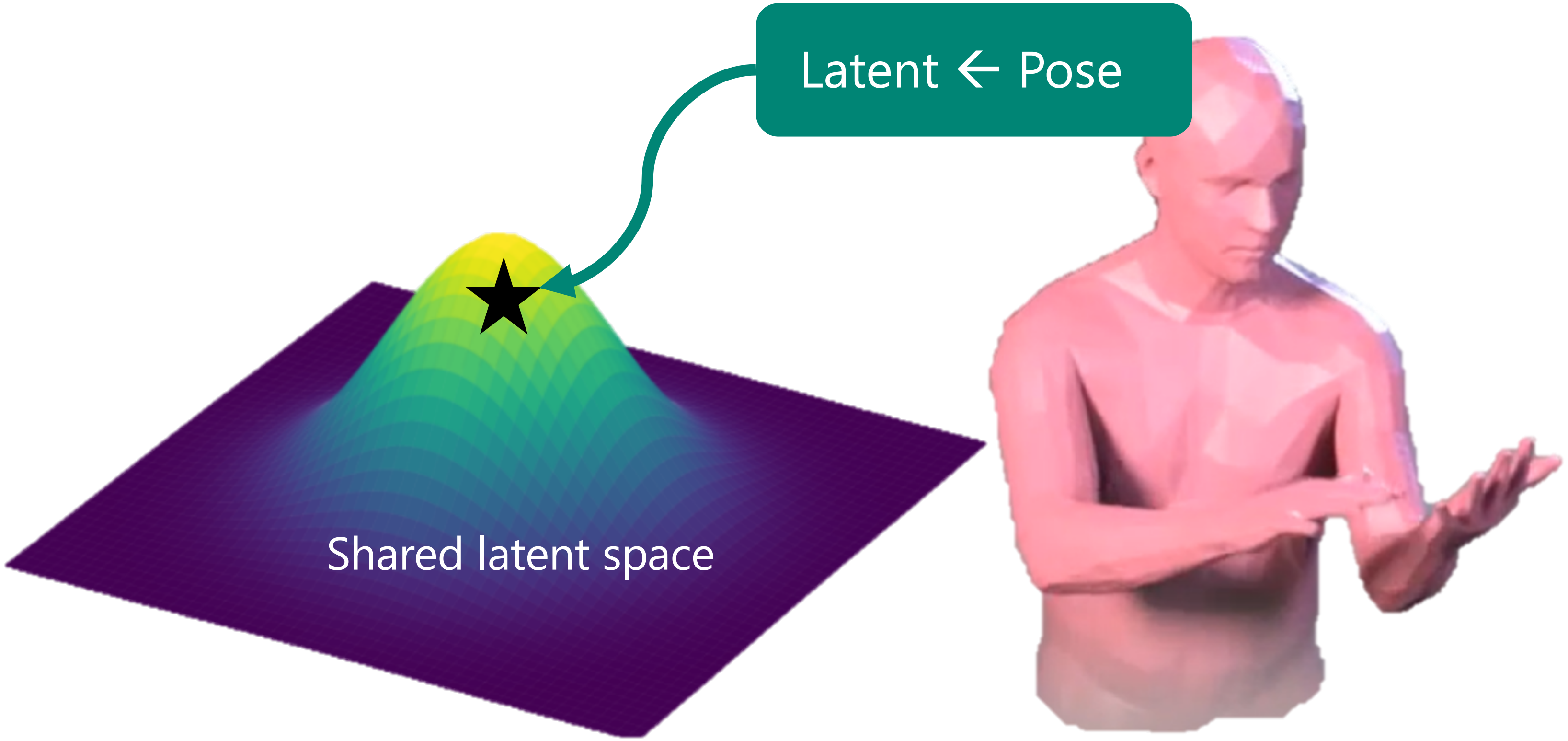
Idea

Latent ← Pose properties

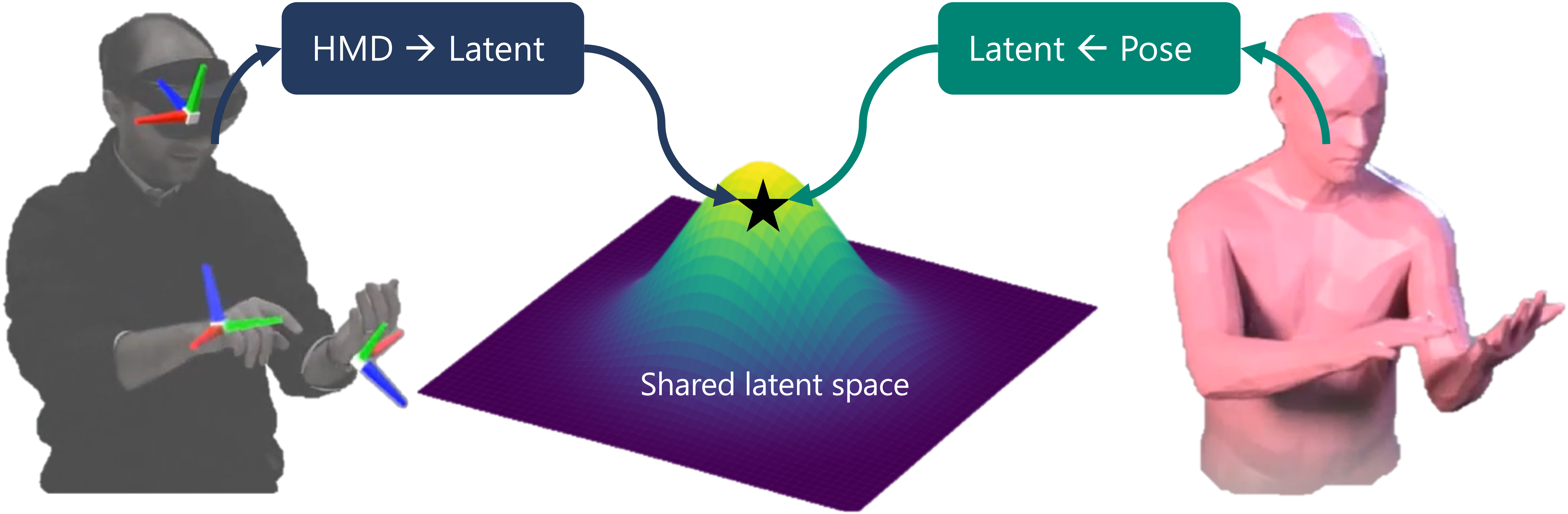
Deal with uncertainty

Expressive enough

Invertibility

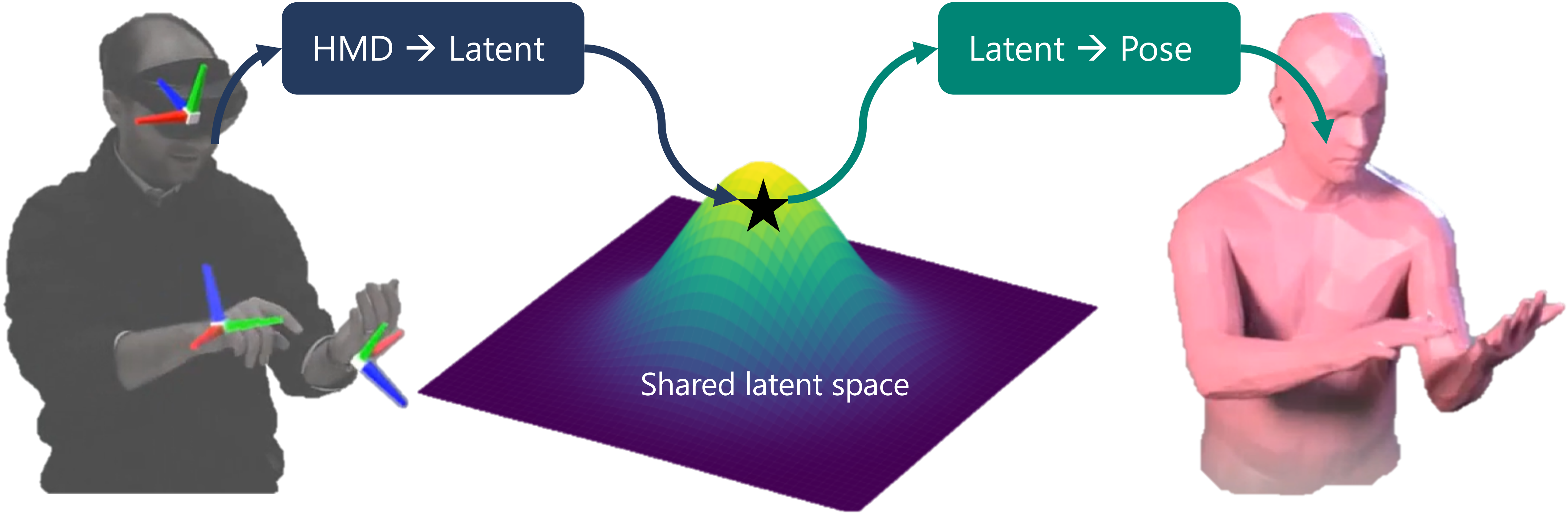


Idea



During training

Idea



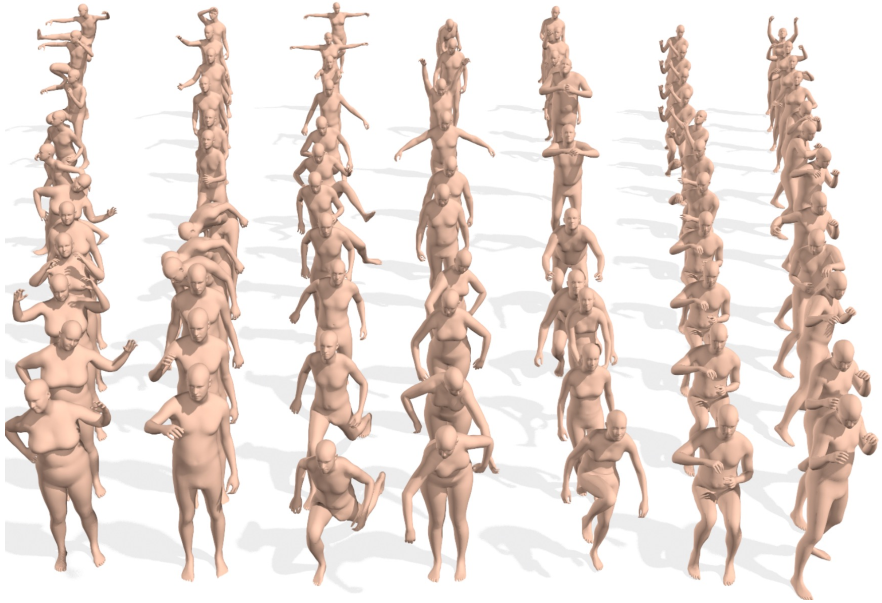
During generation

We call our approach FLAG

Flow-based 3D **A**vatar **G**eneration

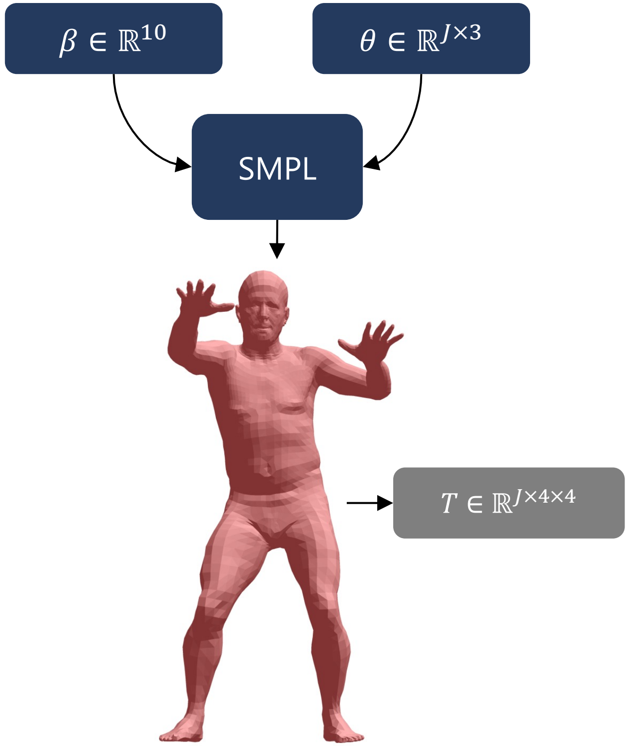
Data

AMASS Dataset



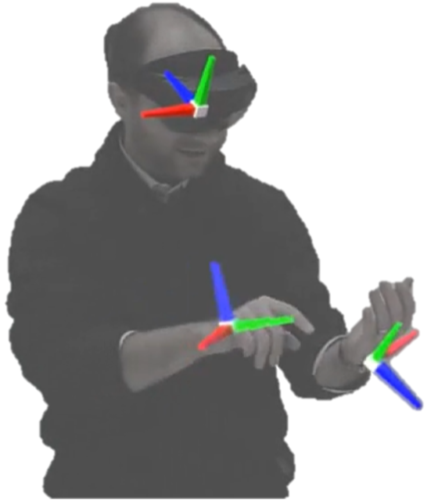
[AMASS (mpg.de)]

SMPL



[SMPL (mpg.de)]

HMD Signal



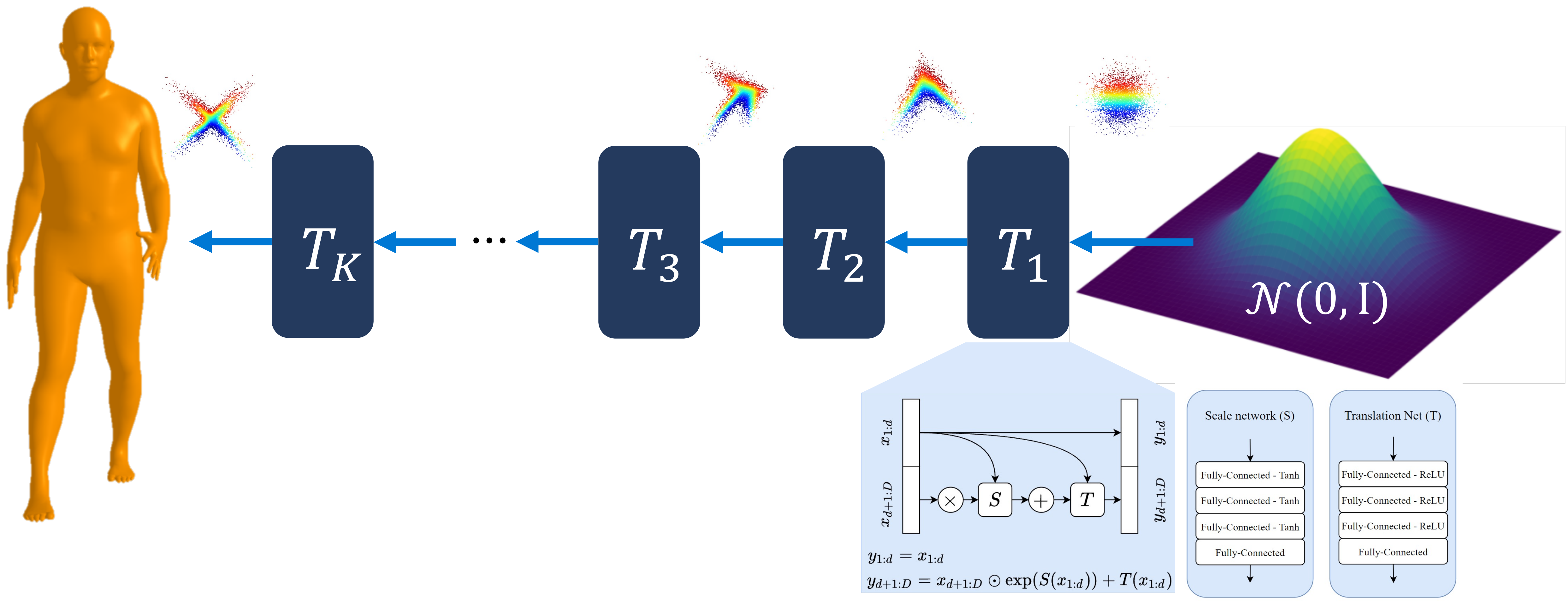
$\mathbb{H} \in \mathbb{R}^{3\times 4\times 4}$

Head rotation and translation
Hands rotation and translation

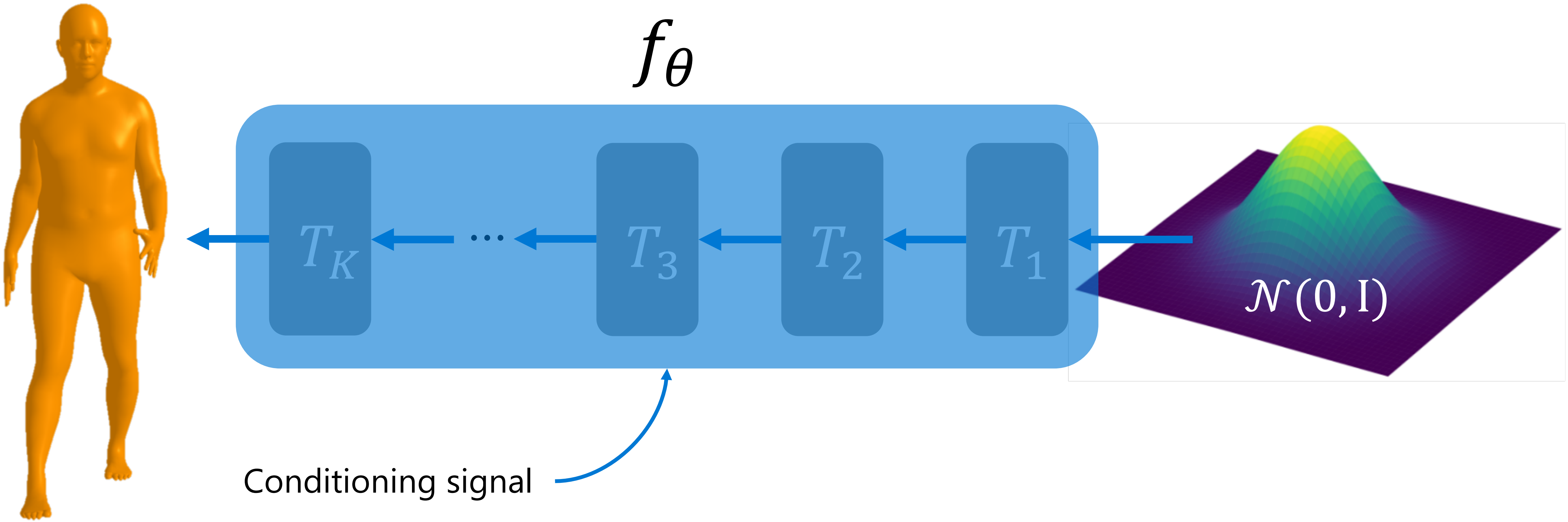
Mapping between 3D pose and the Latent Space

Via a conditional Flow-based model

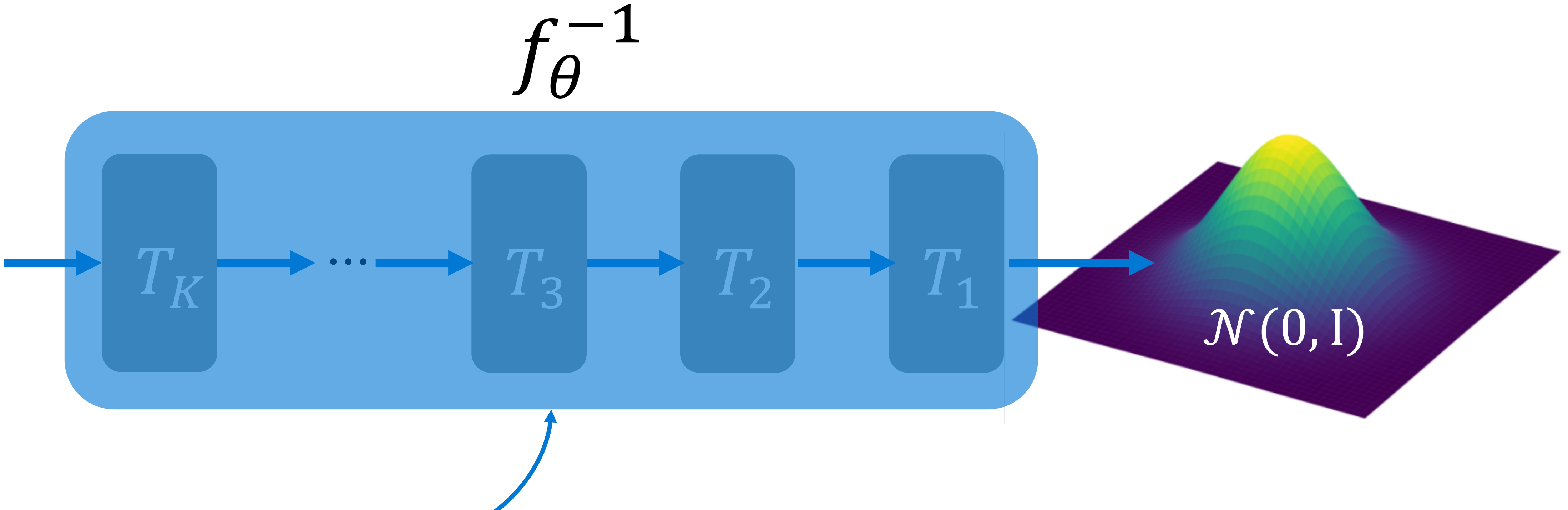
f_θ: A Conditional Flow-based Model



f_θ: A Conditional Flow-based Model



f_θ: A Conditional Flow-based Model

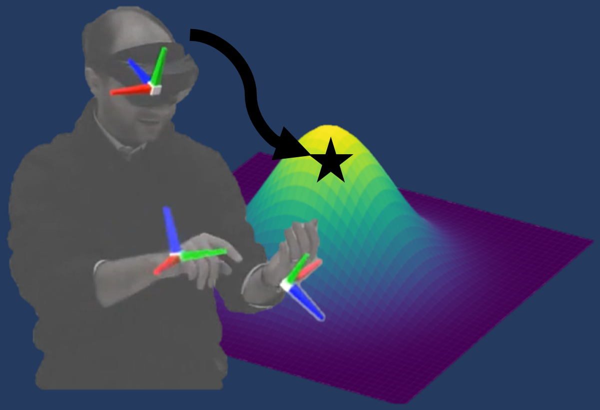


Conditioning signal

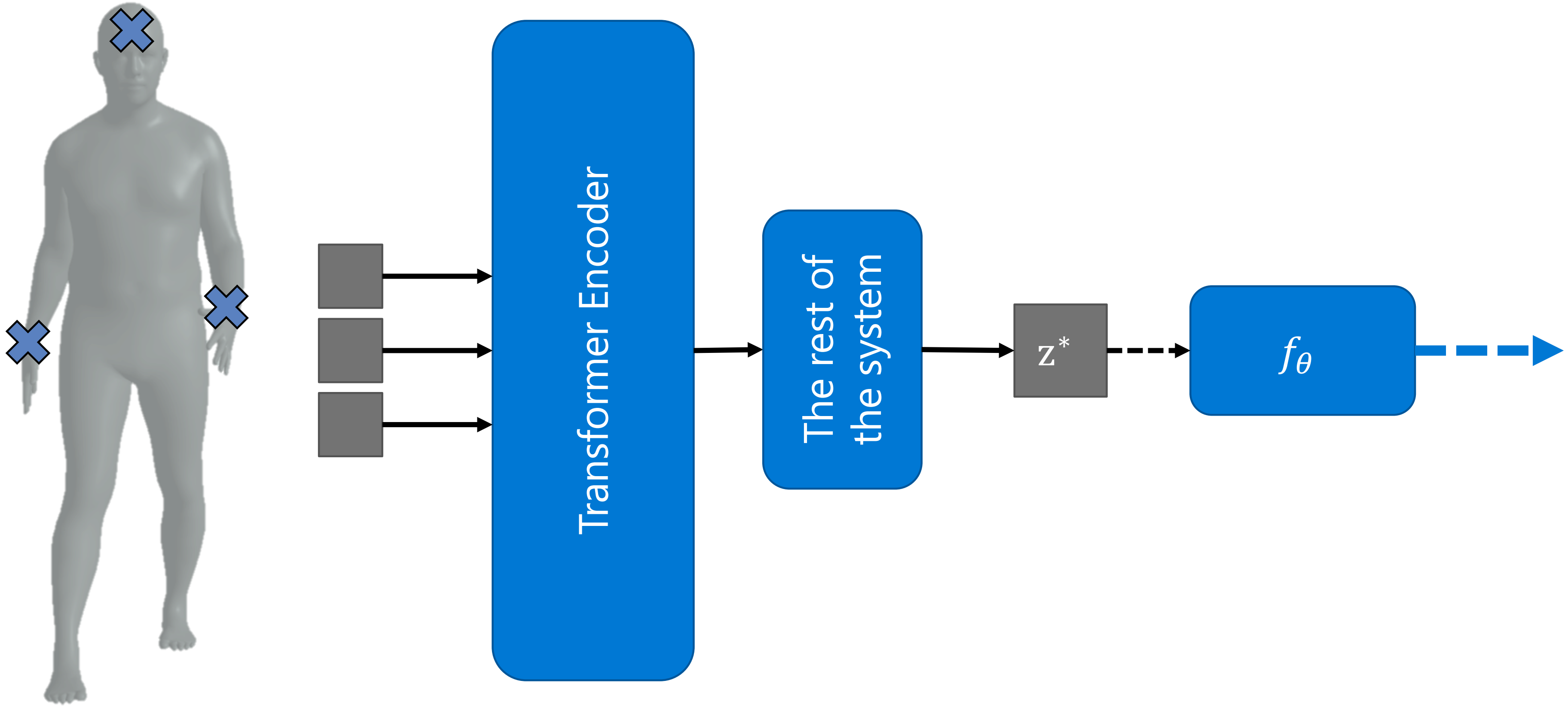
$$\log p(x) = \log p(z_0) - \sum_{i=1}^{K} \log \det \left| \frac{\partial T_i}{\partial z_i} \right|$$

Mapping from HMD to the Latent Space

Via transformers with a discrete latent space

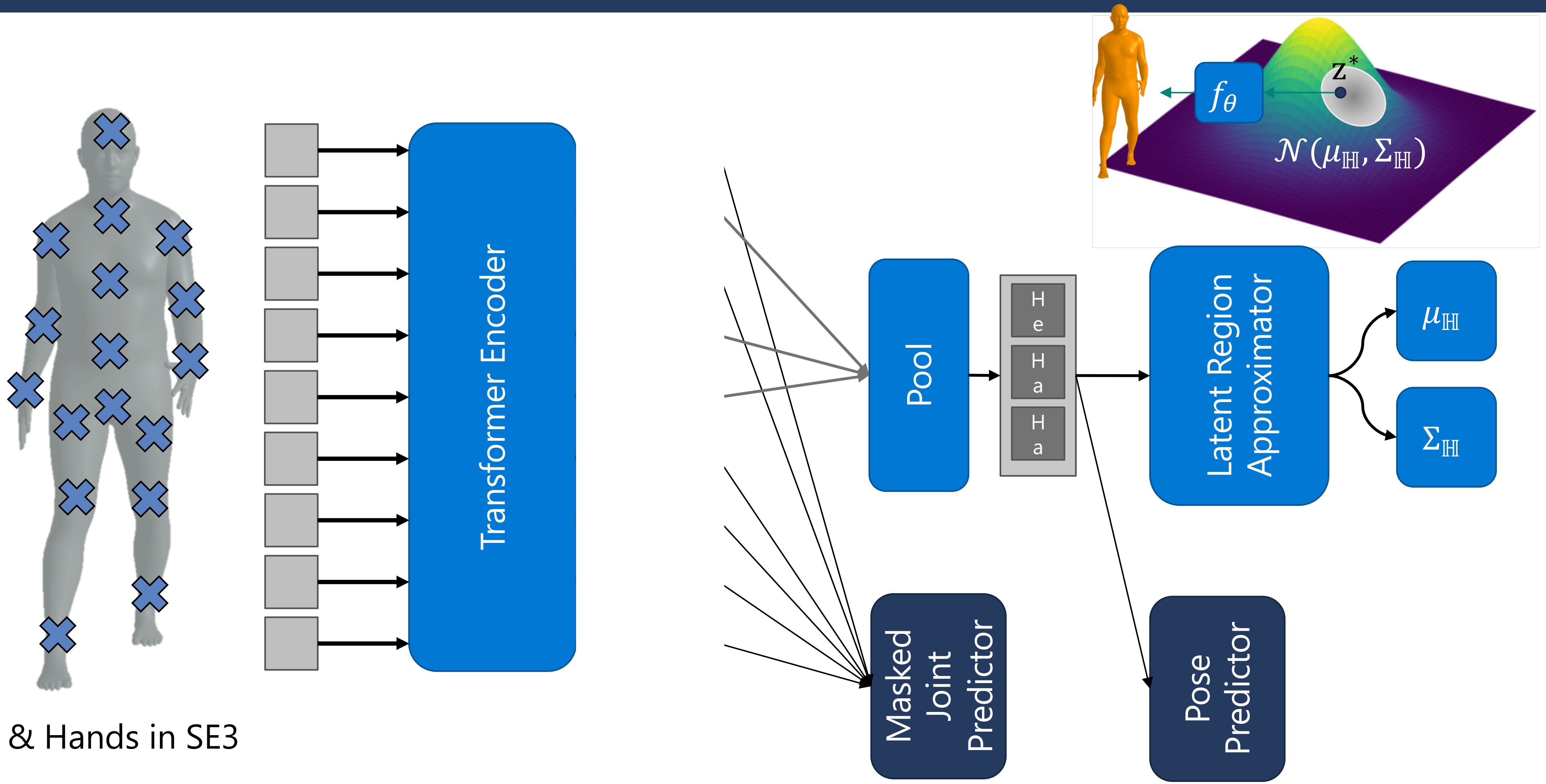


f_{LRA}: Head & Hands → the Latent Space



Head & Hands in SE3

f_{LRA}: Head & Hands → the Latent Space



Head & Hands in SE3

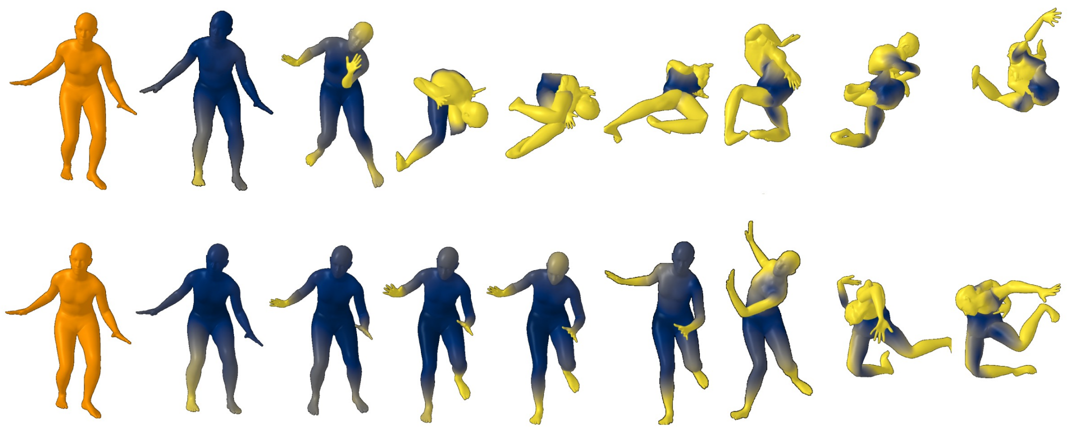
Training and Generation

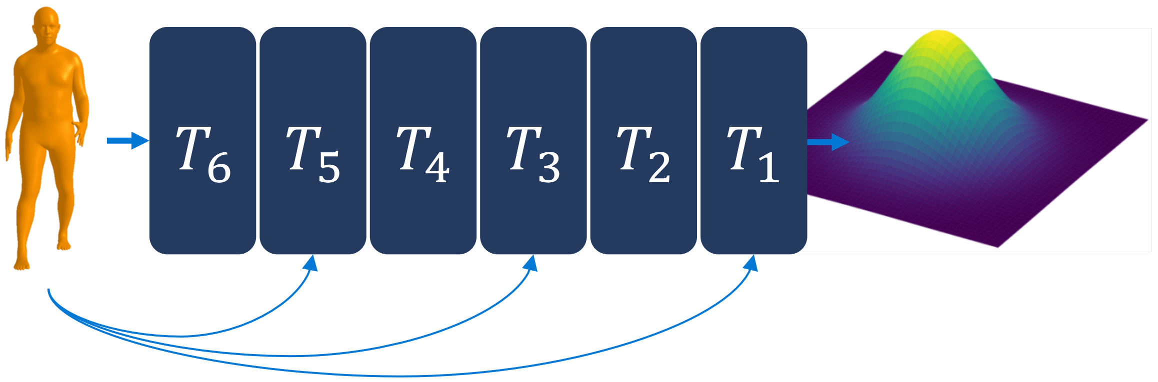
Training

To train f_θ

$$\log p(x) = \log p(z_0) - \sum_{i=1}^{K} \log \det \left| \frac{\partial T_i}{\partial z_i} \right|$$

$$\mathcal{L}_{\text{nll}} = - \log p_\theta(x_\theta)$$





Training

To train f_{LRA}

Auxiliary tasks

$$\mathcal{L}_{\text{mjp}} = \sum_{j \in J_{\text{masked}}} \left\| \hat{x}_P^j - x_P^j \right\|_2^2$$

$$\mathcal{L}_{\text{rec}} = \left\| \hat{x}_\theta^{\text{tps}} - x_\theta \right\|_2^2$$

Latent region approximation

$$\begin{aligned}\mathcal{L}_{\text{lra}} = -\alpha_{\text{nll}} \log p_{\mathbb{H}}(z^*) + \alpha_{\text{rec}} \left\| \mu_{\mathbb{H}} - z^* \right\|_2^2 \\ -\alpha_{\text{reg}}(1 + \ln \sigma_{\mathbb{H}} - \sigma_{\mathbb{H}})\end{aligned}$$

Generation

Direct Prediction

Head & Hands

f_{LRA}

$z = \mu_H$

f_θ

$\hat{\theta}$

Optimization in pose space

Head & Hands

θ_{init}

optimize

θ_{opt}

Optimization in latent space

Head & Hands

f_{LRA}

$z_{init} = \mu_H$

optimize

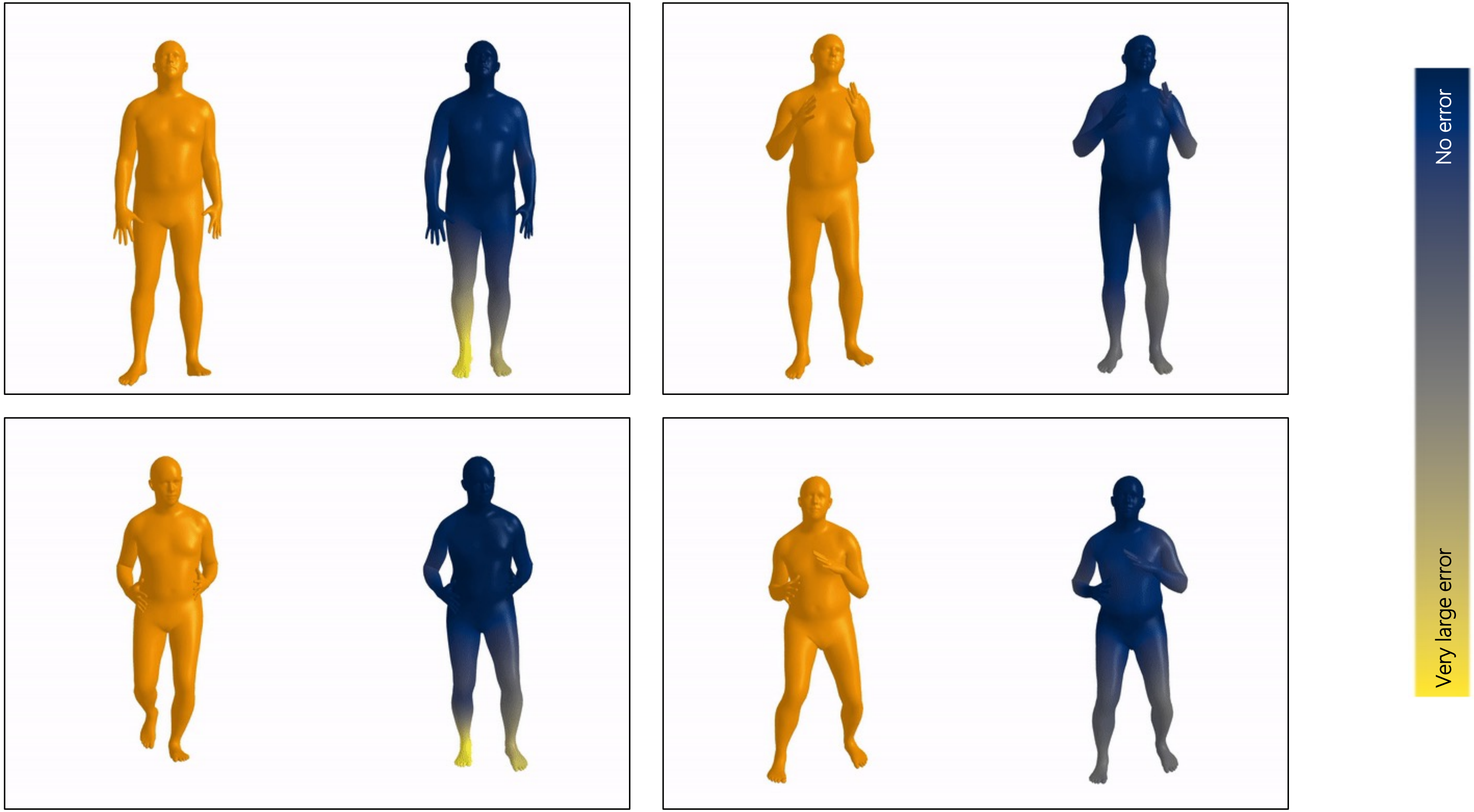
z_{opt}

f_θ

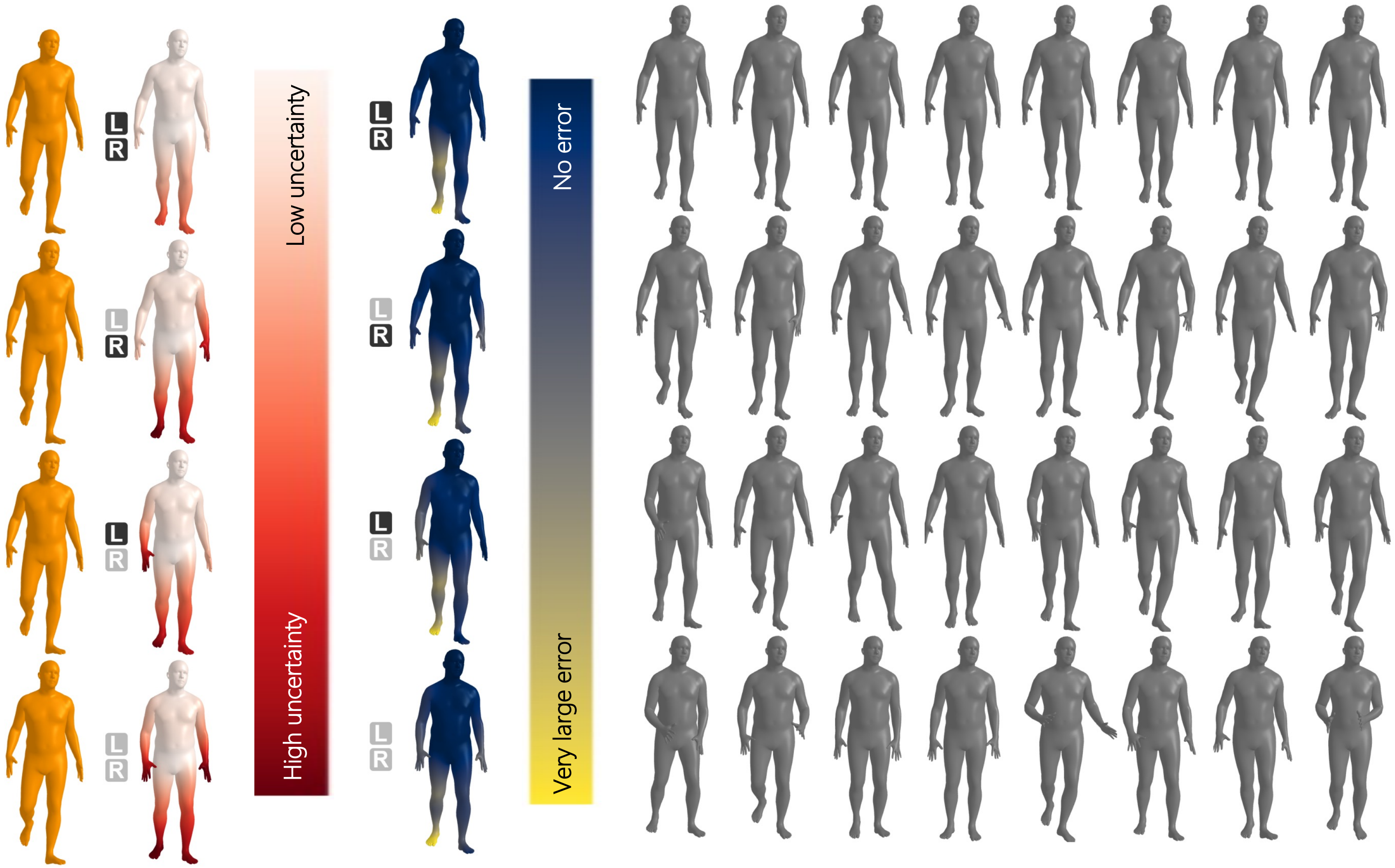
$\hat{\theta}$

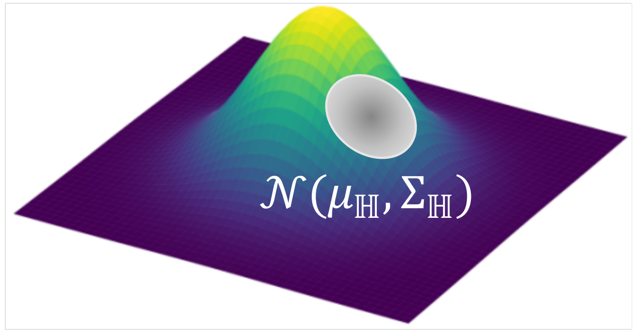
Experiments and Results

Qualitative Results



Qualitative Results



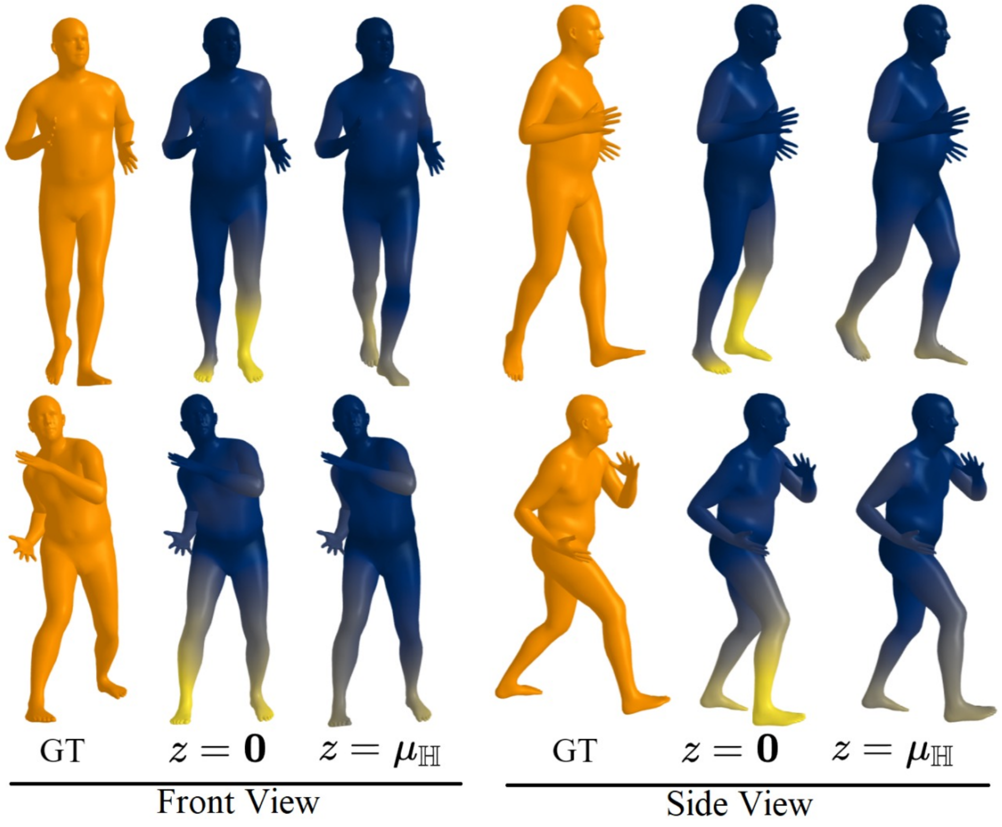


Quantitative Results

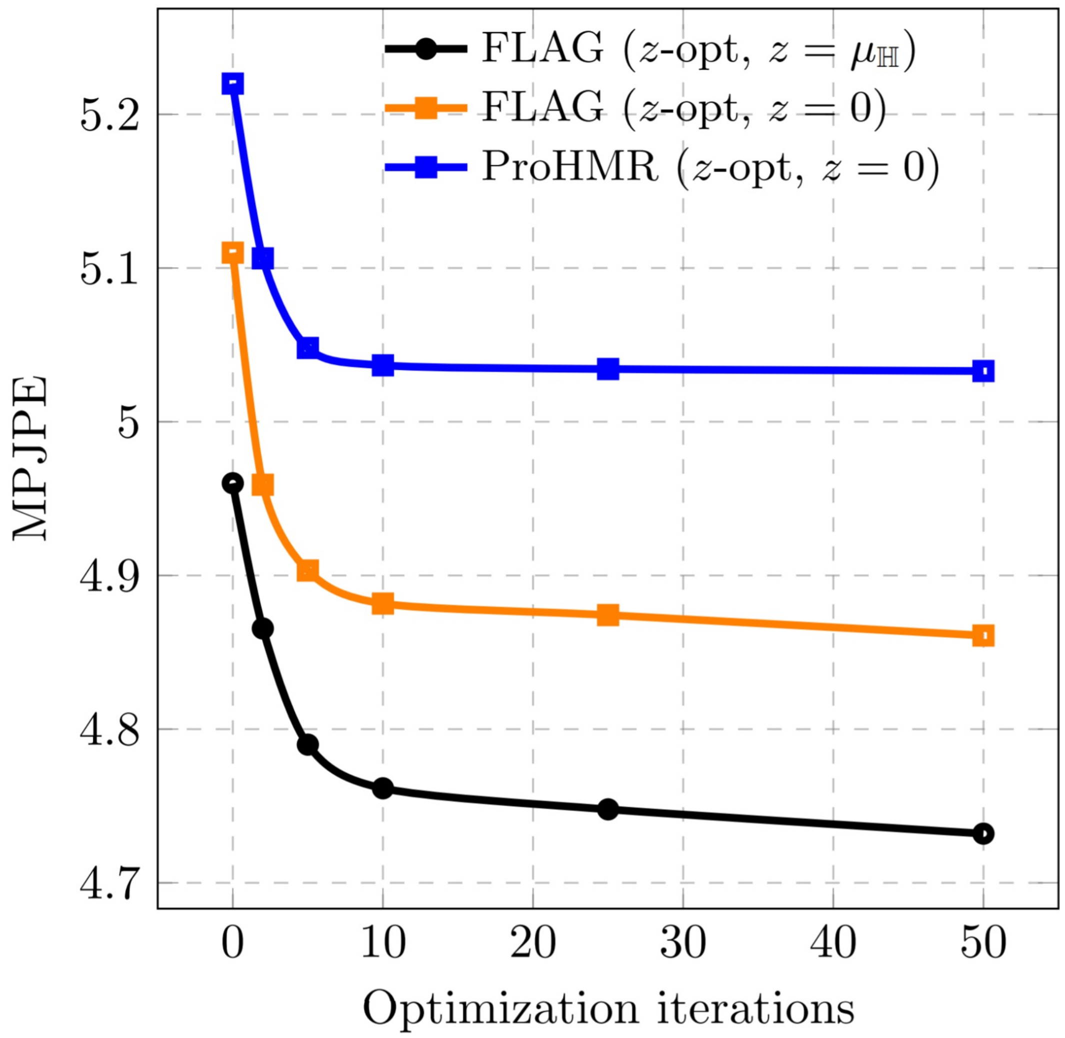
MPJPE: mean per-joint position error

Method	Upper Body MPJPE (↓)	Full Body MPJPE (↓)
VPoser-HMD	1.69 cm	6.74 cm
HuMoR-HMD	1.52 cm	5.50 cm
VAE-HMD	3.75 cm	7.45 cm
ProHMR-HMD	1.64 cm	5.22 cm
FLAG (Ours)	**1.29 cm**	**4.96 cm**

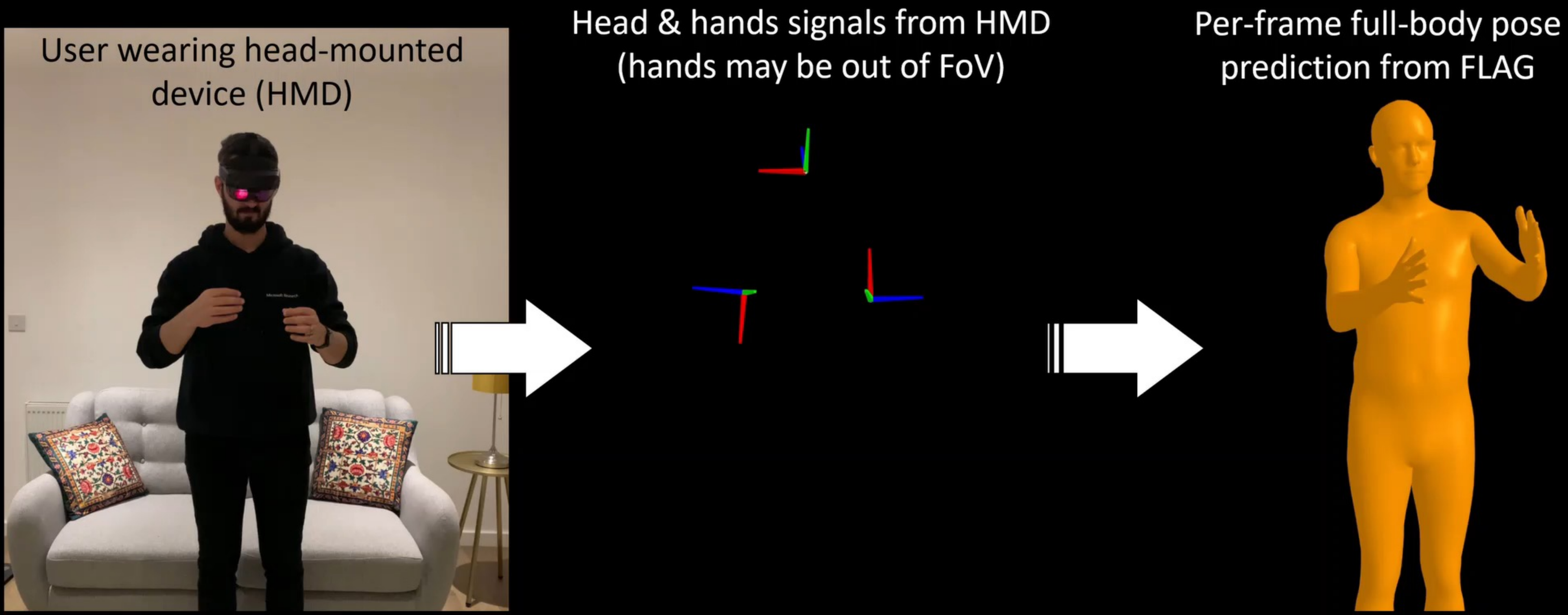
Latent Variable Sampling	Upper Body MPJPE (↓)	Full Body MPJPE (↓)
Zeros ($z = \mathbf{0}$)	1.39 cm	5.11 cm
MLP ($z = \mathrm{MLP}_{\mathbb{H}}$)	1.36 cm	5.05 cm
Ours ($z = \mu_{\mathbb{H}}$)	**1.29 cm**	**4.96 cm**



Optimization



FLAG on HoloLens 2



Microsoft

Conclusion

- *People* are at the heart of mixed reality applications, and so generating realistic human representations with high fidelity is key to the user experience.
- FLAG presents a solution to the extremely challenging problem of generating realistic and high-fidelity full-body poses from sparse HMD observations

Thank you

https://microsoft.github.io/flag