MEGA: Masked generative autoencoder for human mesh recovery

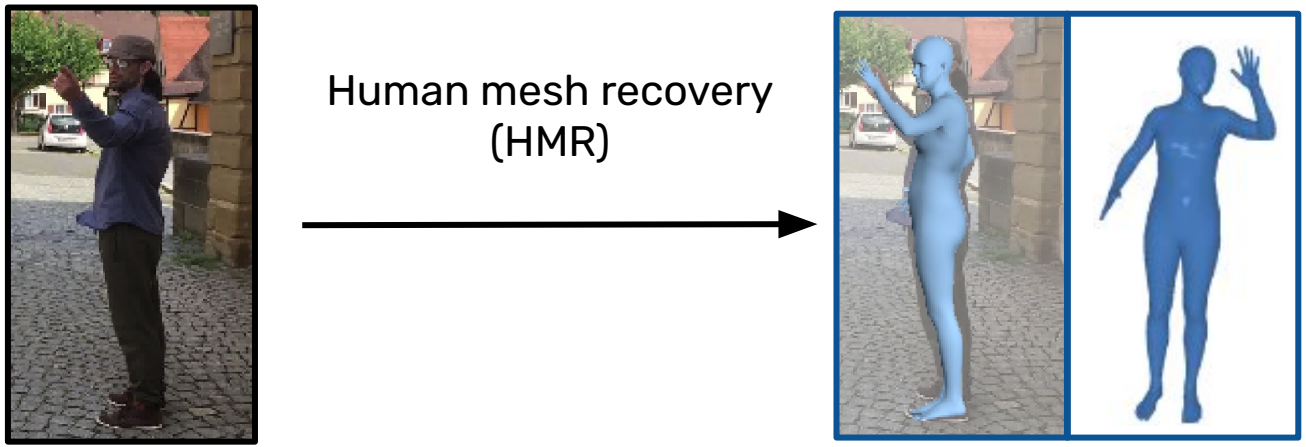
Guénolé Fiche[1,2] Simon Leglaive[1] Xavier Alameda-Pineda[3] Francesc Moreno-Noguer[4,5]

[1]CentraleSupélec, IETR [2]Naver Labs Europe [3]Inria, Univ. Grenoble Alpes
[4]Institut de Robòtica i Informàtica Industrial [5]Amazon

CVPR Nashville JUNE 11-15, 2025

Human 3D mesh recovery

Given an image, the task is to recover a human 3D mesh following the SMPL[1] topology.



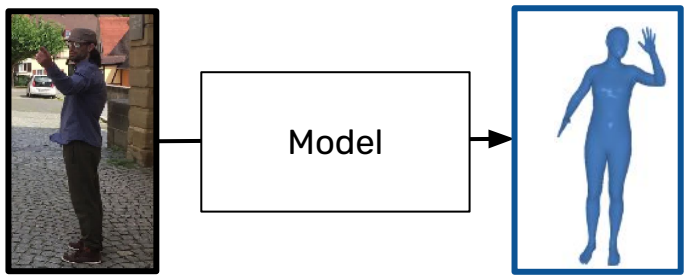
1. Loper, Matthew, et al. "SMPL: A Skinned Multi-Person Linear Model." *ACM Transactions on Graphics* 34.6 (2015).

An ambiguous problem

Many **different 3D meshes yield the same 2D projection** due to occlusions and depth ambiguity.

Single and multi-output methods propose different ways of addressing this issue.

Single output methods

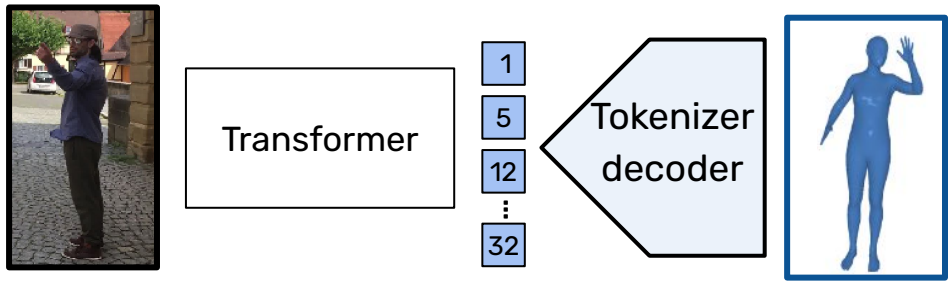


- Give the most likely output
- Efficiency and accuracy
- TokenHMR[2] (CVPR'24)
- VQ-HPS[3] (ECCV'24)

2. Dwivedi, Sai Kumar, et al. "TokenHMR: Advancing human mesh recovery with a tokenized pose representation." CVPR, 2024.
3. Fiche, Guénolé, et al. "VQ-HPS: Human pose and shape estimation in a vector-quantized latent space." ECCV, 2024.

Single output methods - Observation

These recent SOTA approaches use **Transformer-based architectures**, and rely on a **tokenized representation** of human meshes.



Single

Wishlist

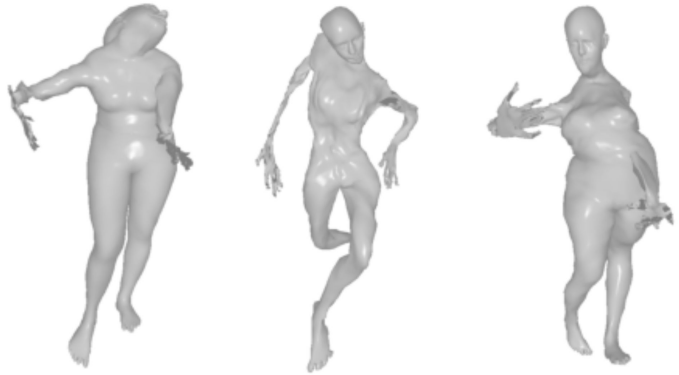
- **Tokenized latent space:** implicit prior on human meshes

These

n a

Single output methods - Weaknesses

Despite tokenization, some sequences may correspond to **unrealistic human meshes**. Additionally, the **diversity** of meshes is **very limited** in HMR training sets.



Meshes obtained with purely random indices.

Single

Despit
Addit

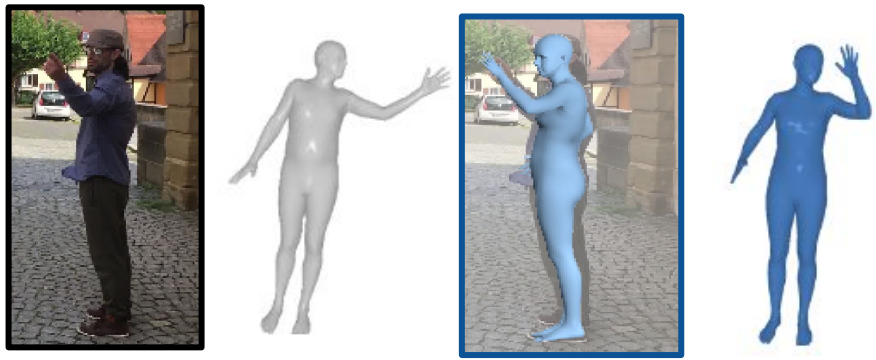
nes.
ts.

Wishlist

- Tokenized latent space
- **Motion capture pre-training:** prior on token sequences

Single output methods - Weaknesses

Overall, single output methods **overlook the ambiguity issue**. In some scenarios, the prediction may differ significantly from the ground truth.



Image, ground truth, and prediction of a single output model.

Single

Overa

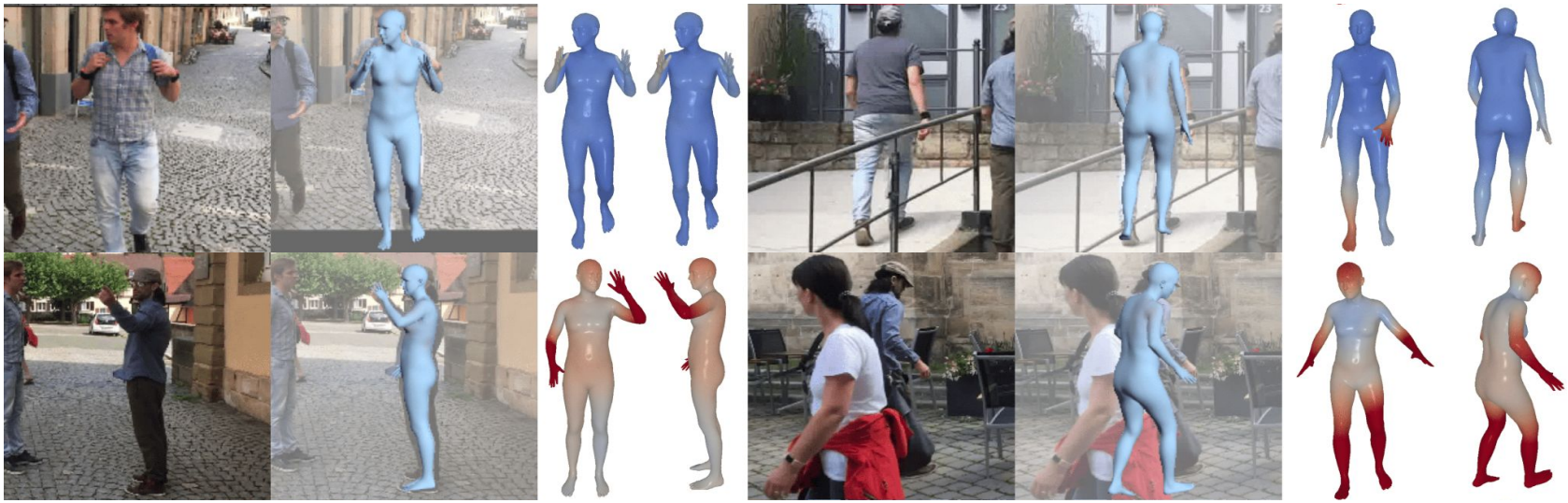
he

Wishlist

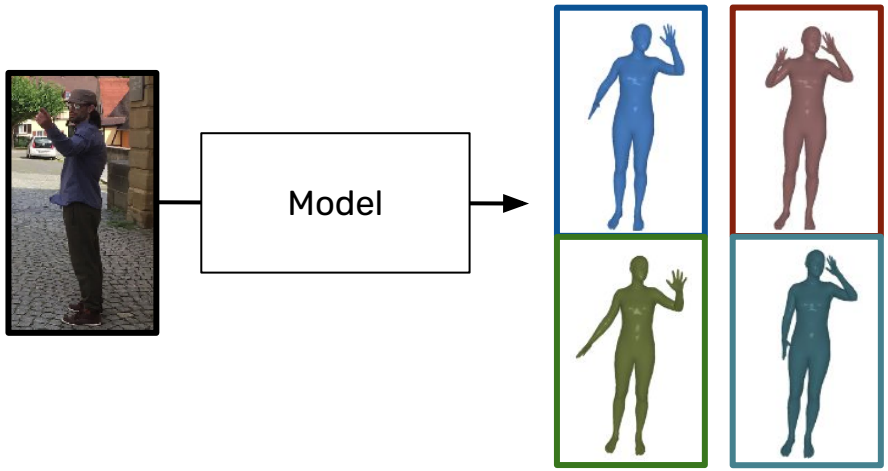
- ❏ Tokenized latent space
- ❏ Motion capture pre-training
- ❏ **Generative model:** multi-output HMR

How do multi-output methods address ambiguity?

We predict multiple meshes per image using our method. The **standard deviation** of the vertices position can be interpreted as **a measure of per-vertex uncertainty, highly related to ambiguity.**



Multi-output methods

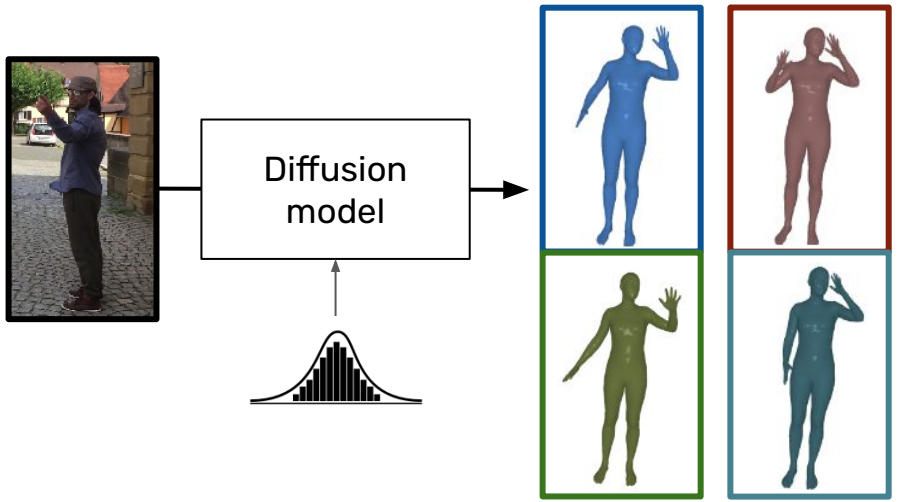


- Set of potential solutions
- Address the ambiguity
- Diff-HMR[4] (ICCV'23)
- ScoreHypo[5] (CVPR'24)

4. Cho, Hanbyel, and Junmo Kim. "Generative approach for probabilistic human mesh recovery using diffusion models." ICCV, 2023.
5. Xu, Yuan, et al. "ScoreHypo: Probabilistic human mesh estimation with hypothesis scoring." CVPR, 2024.

Multi-output methods - Weaknesses

These recent SOTA multi-output methods are based on diffusion models. Those models are powerful but have a high computational cost.



Multi-o

The

Wishlist

- ❑ Tokenized latent space
- ❑ Motion capture pre-training
- ❑ Generative model
- ❑ **Efficient, fast prediction**

Multi-output methods - Weaknesses

When making a single prediction, multi-output methods **lack accuracy** because giving a zero noise as an input does not necessarily give the most likely solution.

	MPJPE on 3DPW (in mm)
Diff-HMR (single prediction)	98.9
VQ-HPS (same backbone and training data)	79.1

Multi-o

When

ing

Wishlist

- ❑ Tokenized latent space
- ❑ Motion capture pre-training
- ❑ Generative model
- ❑ Efficient, fast prediction
- ❑ **Single output:** competitive with single output methods

Wishlist

- ❑ Tokenized latent space
- ❑ Motion capture pre-training
- ❑ Generative model
- ❑ Efficient, fast prediction
- ❑ Single output

One approach that will allow us to check all these boxes is **Masked Generative Modeling[6].**

6. Chang, Huiwen, et al. "MaskGIT: Masked generative image transformer." CVPR, 2022.

Overview of the contributions

- NEW Masked generative modeling for HMR
- NEW Unconditional distribution of 3D meshes before conditioning on images
- NEW A single model for 3 different tasks:

Unconditional human mesh generation

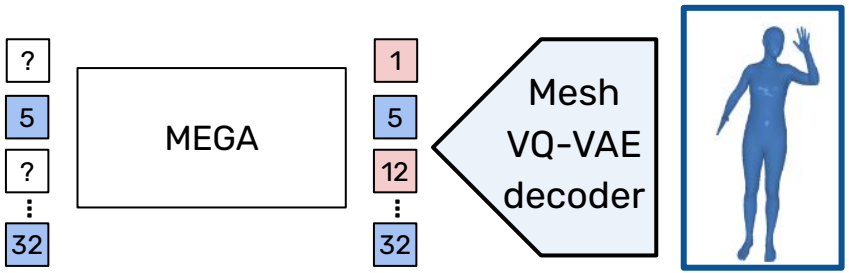
Single output HMR

Multi-output HMR

Self-supervised pre-training on human meshes

Contribution 1
Contribution 2

We rely on **Mesh-VQ-VAE**[3], that represents a human mesh as a **sequence of 54 tokens.** MEGA is **masked autoencoder**[7] in the space of tokenized meshes: it learns to recover human meshes given **partial sequences of tokens.**



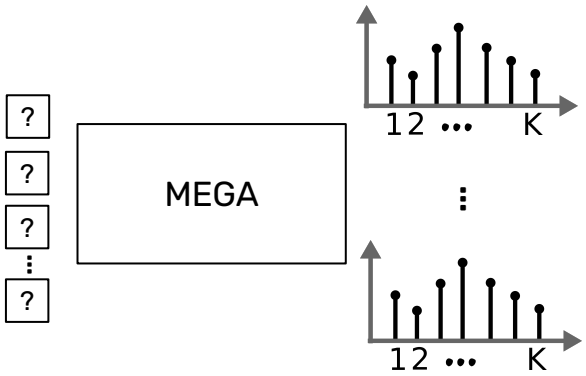
MEGA is trained on AMASS with **a single cross-entropy loss** on the tokenized human meshes.

3. Fiche, Guénolé, et al. "VQ-HPS: Human pose and shape estimation in a vector-quantized latent space." ECCV, 2024.
7. He, Kaiming, et al. "Masked autoencoders are scalable vision learners." CVPR, 2022.

Masked generative modeling on human meshes

Contribution 1
Task 1

Once pre-trained on motion capture data, MEGA can generate **unconditional human meshes** using the **masked generative modeling**[6] strategy.

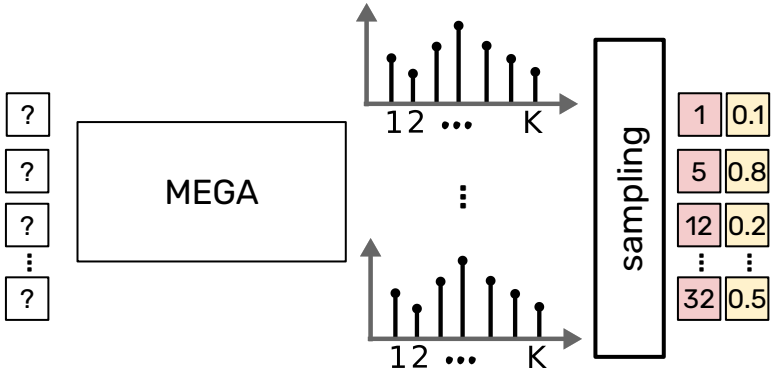


We start from a **fully masked sequence** and predict a **distribution over indices** for each token.

6. Chang, Huiwen, et al. "MaskGIT: Masked generative image transformer." CVPR, 2022.

Masked generative modeling on human meshes

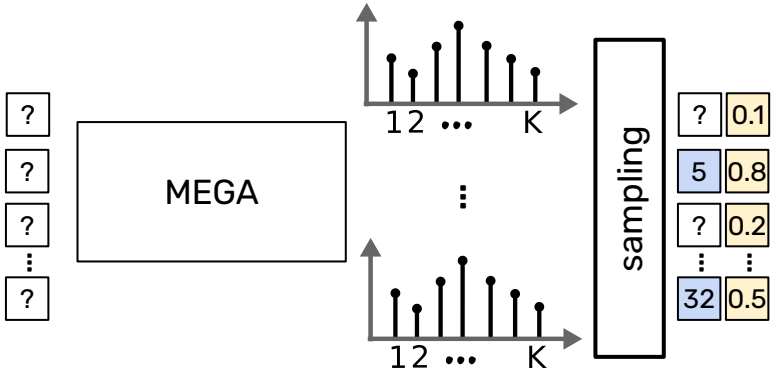
Contribution 1
Task 1



We **sample the distributions**, and **keep the confidence value** associated with the selected tokens.

Masked generative modeling on human meshes

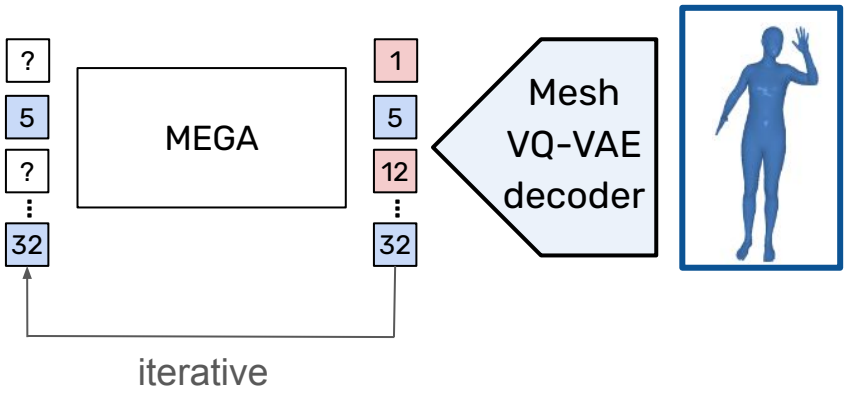
Contribution 1
Task 1



We select the tokens with highest confidence values, **freeze them for the next steps**.
Other tokens are masked and will be predicted in the next steps.

Masked generative modeling on human meshes

Contribution 1
Task 1



We feed the frozen tokens to our model and iterate this process until the sequence is predicted. We can then decode it to obtain the generated mesh.

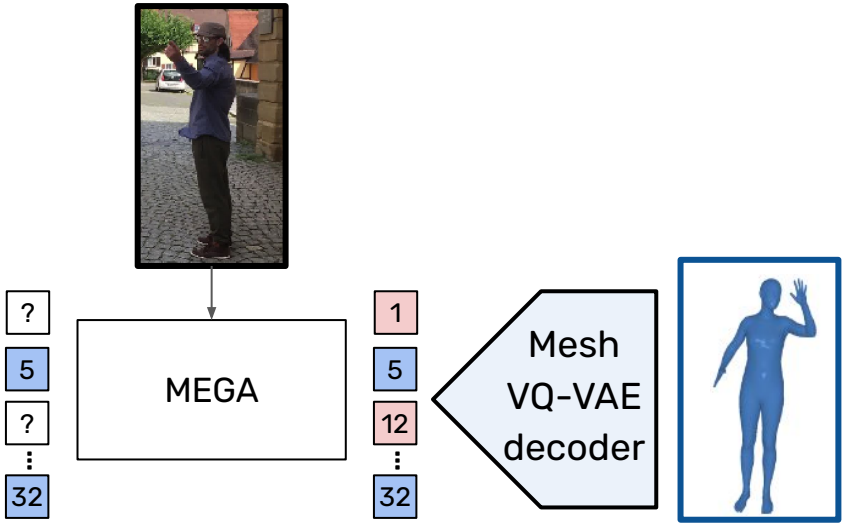
Random human mesh generation

Task 1

Once we learnt the **unconditional distribution** of human meshes, we aim to **condition the generation on an image**.

MEGA - Supervised fine-tuning for HMR

Contribution 1

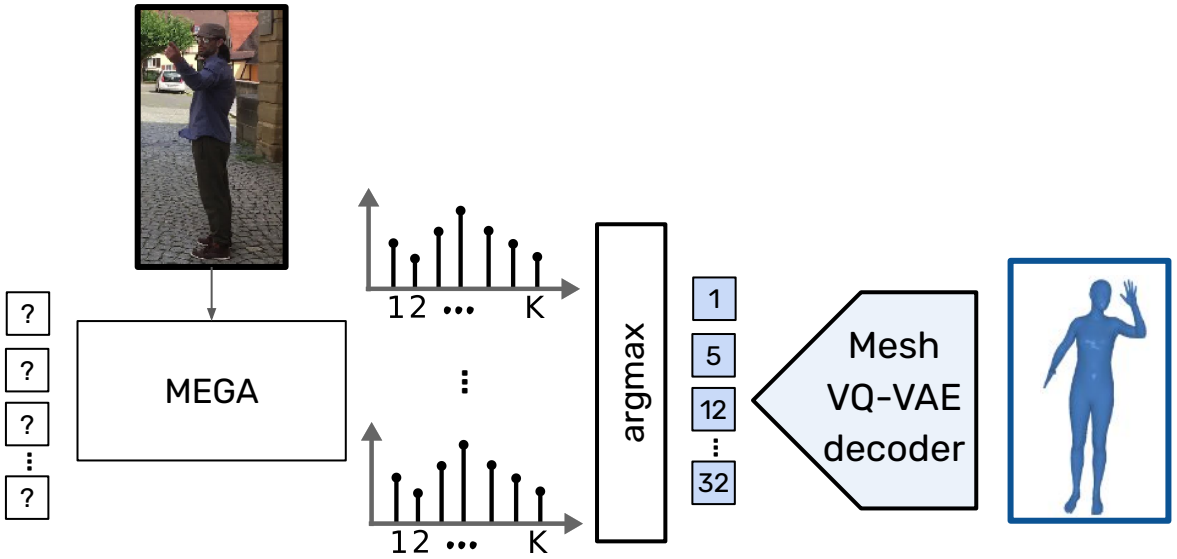


We add image conditioning and learn to recover a human mesh from **an image and a partial sequence of tokens**.

Once again, our **only loss** for predicting human meshes is the cross-entropy.

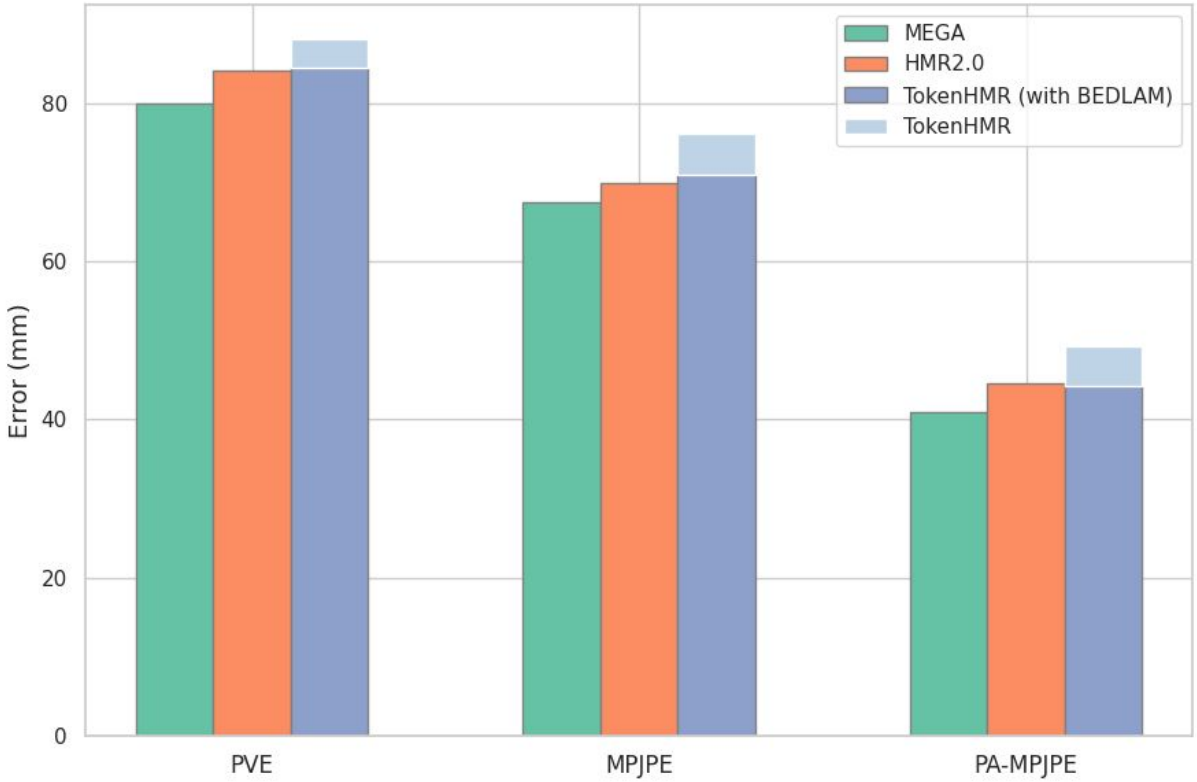
Single output HMR

Task 2



In order to obtain a single prediction, the **tokens are selected with Argmax**. We predict all tokens in a **single forward pass** (0.03 sec. with ResNet).

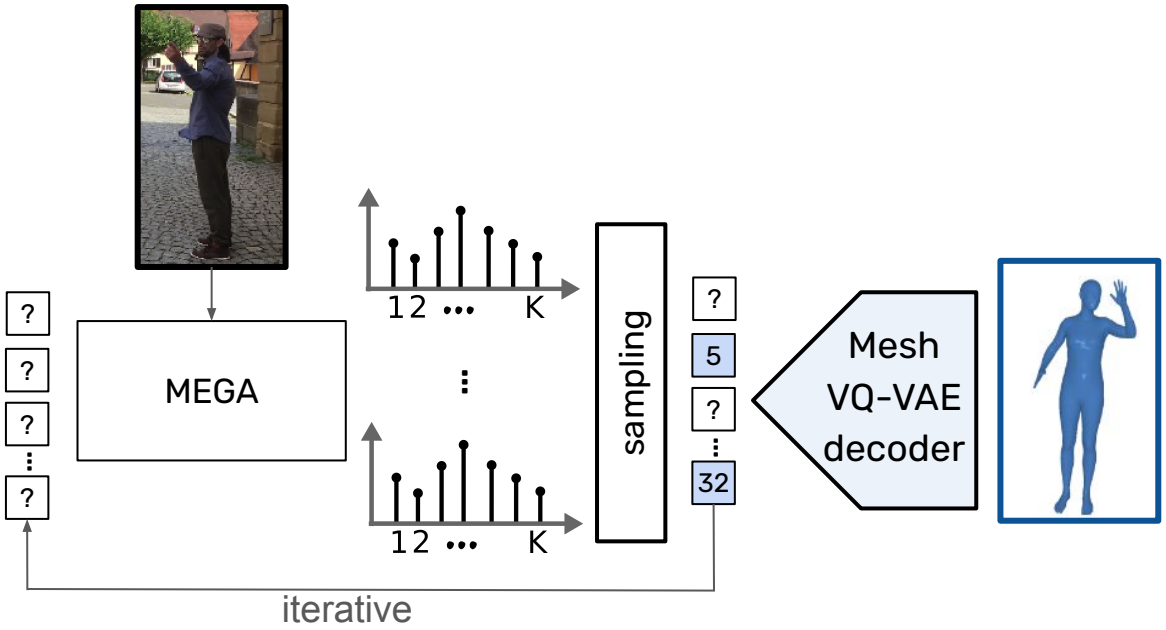
Single output evaluation on 3DPW



2. Dwivedi, Sai Kumar, et al. "Tokenhmr: Advancing human mesh recovery with a tokenized pose representation." CVPR, 2024.
3. Fiche, Guénolé, et al. "VQ-HPS: Human pose and shape estimation in a vector-quantized latent space." ECCV, 2024.
8. Goel, Shubham, et al. "Humans in 4D: Reconstructing and tracking humans with transformers." ICCV, 2023.

Multi-output HMR

Task 3

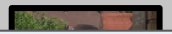


To obtain diverse plausible predictions, we again use the **masked generative modeling sampling scheme.**

Predictions are made in only **5 iterations** (0.04 sec. for 16 samples).

Multi-output HMR

Task 3



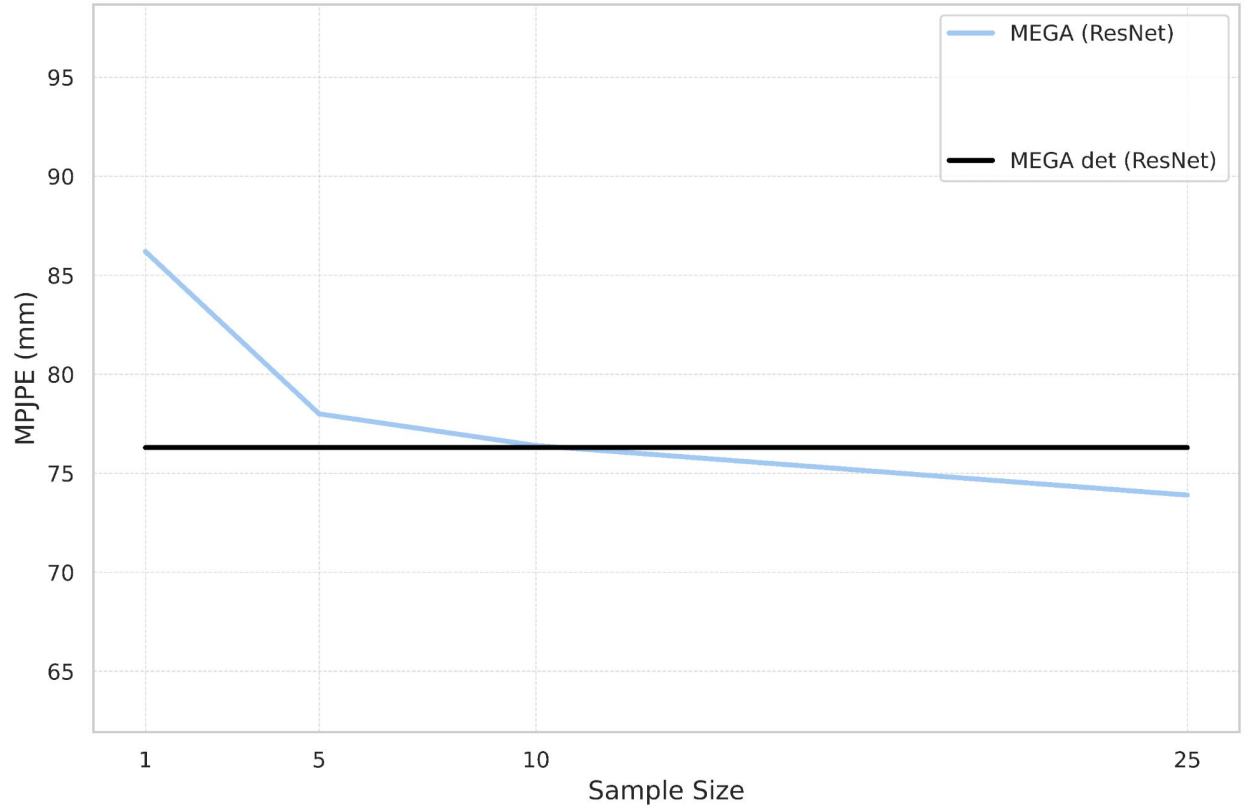
Wishlist

- ✓ Tokenized latent space
- ✓ Motion capture pre-training
- ✓ Generative model
- ✓ Efficient, fast prediction
- ✓ Single output

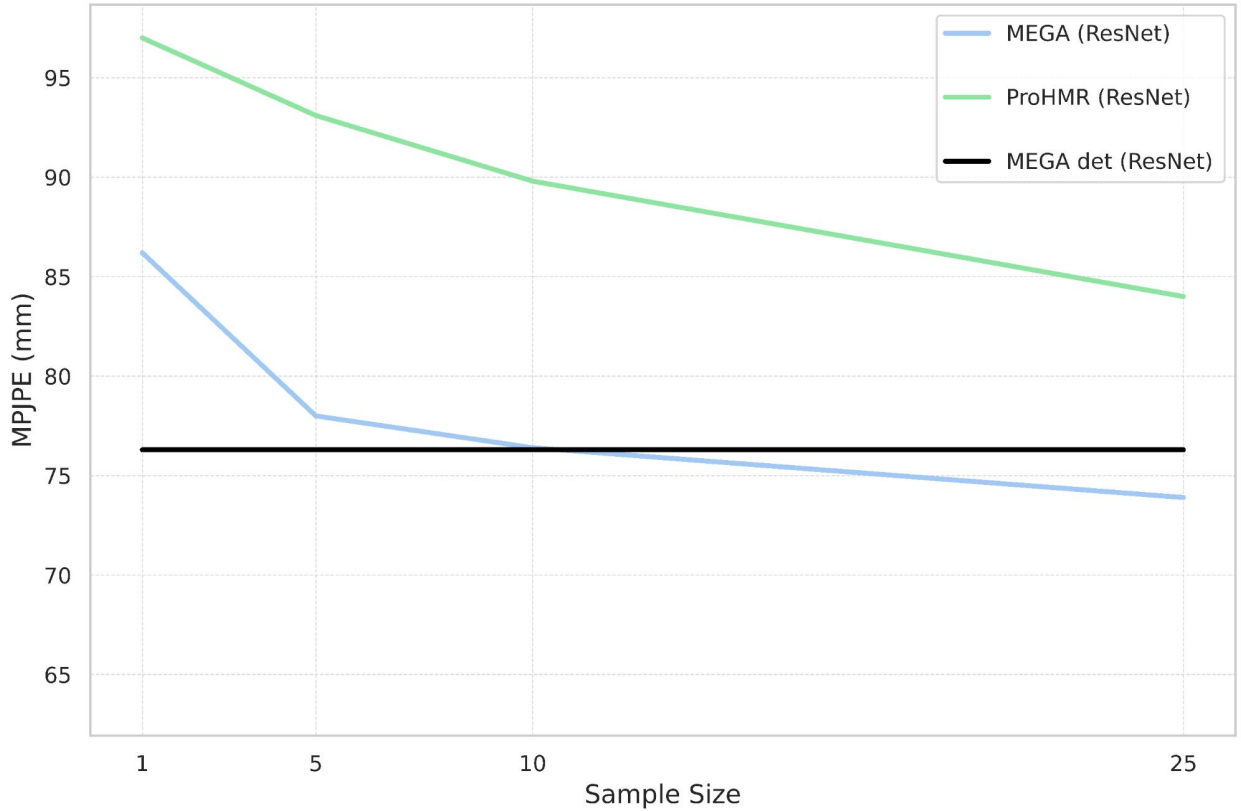
To obtain diverse plausible predictions, we again use the masked generative modeling sampling scheme.

Predictions are made in only 5 iterations (0.04 sec. for 16 samples).

Multi-output evaluation

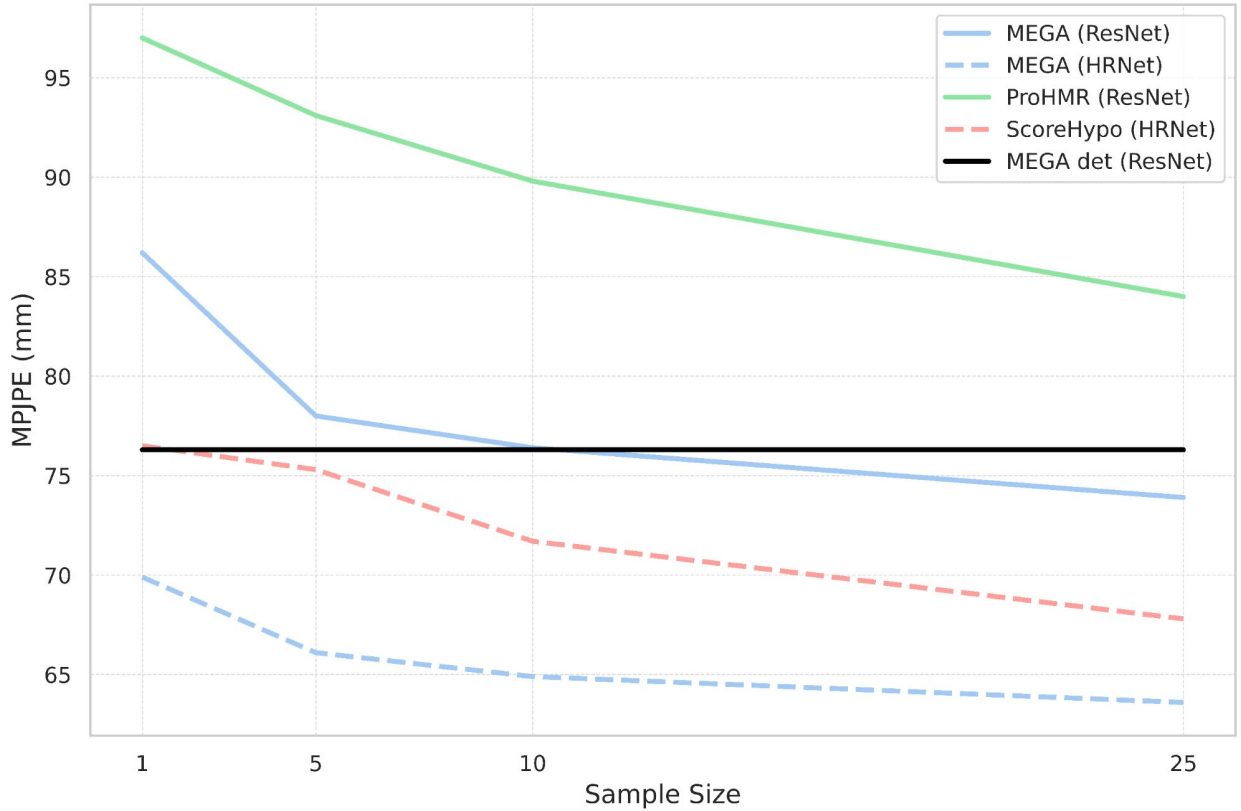


Multi-output evaluation



9. Kolotouros, Nikos, et al. "Probabilistic modeling for human mesh recovery." ICCV, 2021.

Multi-output evaluation



5. Xu, Yuan, et al. "ScoreHypo: Probabilistic human mesh estimation with hypothesis scoring." CVPR, 2024.

Why is masked generative modeling adapted to HMR?

$n_t = 2$

In the first iteration, very few tokens with high confidence are predicted. The prediction is still close to the mean pose.

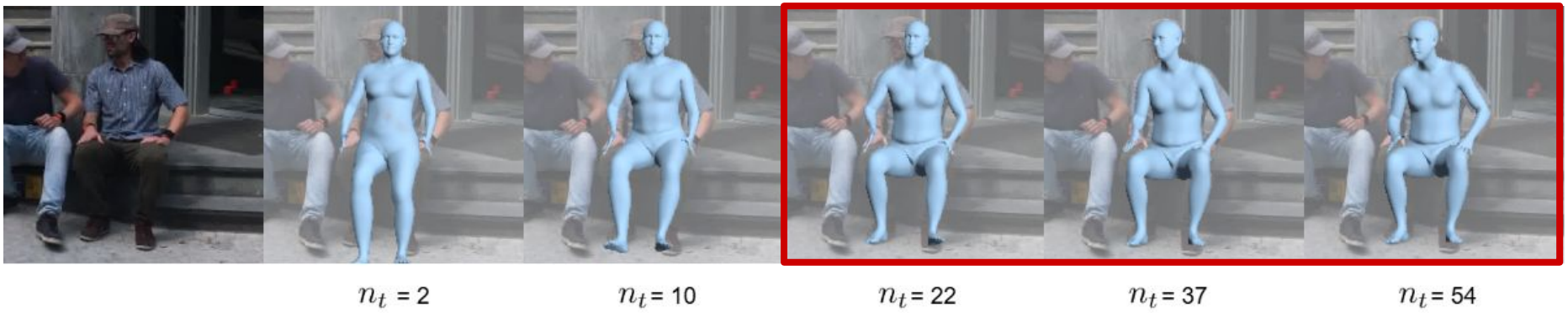
Why is masked generative modeling adapted to HMR?

$n_t = 2$ $n_t = 10$

With only 10 predicted tokens, we start having a **coarse prediction**.

This is because the **tokens with high influence on the pose are easier to predict**, and are then fixed in the first steps.

Why is masked generative modeling adapted to HMR?



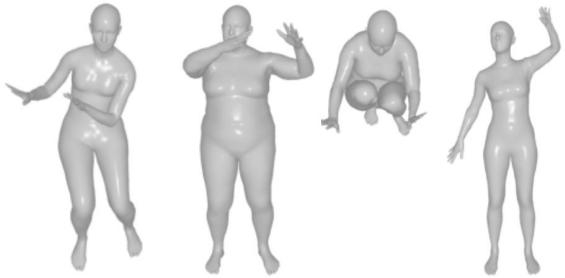
In the last 3 iterations, the mesh becomes more and more **plausible and aligned with the image.**

The last predicted tokens correspond to **fine-grained details**, which are harder to predict.

Key takeaways

MEGA is the first HMR approach based on **masked generative modeling**. It is trained with a **single cross-entropy loss** and can leverage human meshes **with or without images**.

Our **flexible** framework can be used for **diverse applications**:



Overall, MEGA is a step towards **foundation models for 3D humans**.

Thank you for your attention!

https://github.com/g-fiche/MEGA

https://g-fiche.github.io/research-pages/mega/

guenole.fiche@naverlabs.com

Poster #90

DIGISPORT GRADUATE SCHOOL

This study is part of the EUR DIGISPORT project supported by the ANR within the framework of the PIA France 2030 (ANR-18-EURE-0022). This work was performed using HPC resources from the “Mésocentre” computing center of CentraleSupélec, ENS Paris-Saclay, and Université Paris-Saclay supported by CNRS and Région Île-de-France. We would like to thank Samir Sadok for his valuable help with the implementation of the masked autoencoder.