

ICMV, 2026

3D Human Reconstruction using Vision Language Models via Quantized Tokens

Gulraiz Khan, Maneeha Rani, Waqas Ahmed, Kevin Pimblet, Kenneth Y. Wertheim

Data Science, AI and Modelling Centre (DAIM), University of Hull, UK

G.Khan-2022@hull.ac.uk



What is 3D Human Reconstruction?

- Converting 2D image to 3D mesh
- Input a flat 2D photograph of a subject
- Estimate pose, body shape, and position
- Predict mesh or parameter values (SMPL-X)
- Recreate a fully 3D digital mesh model of the subject

Applications of 3D Human Reconstruction?



VR/AR



Film



Gaming



Driving



Health



Forensics

Why is 3D reconstruction hard?



1

Scene occlusion

Table hides his legs

2

Self-occlusion

Her face and hands are hidden

3

Depth & scale

Two people, two distances

4

Head-mounted camera

Moving, downward view

What is SMPL-X?

122 parametric representation of a full 3D body mesh



10

Shape β

Height and build

99

Pose θ

Rotations of body, hands, jaw, eyes

10

Expression ψ

Face

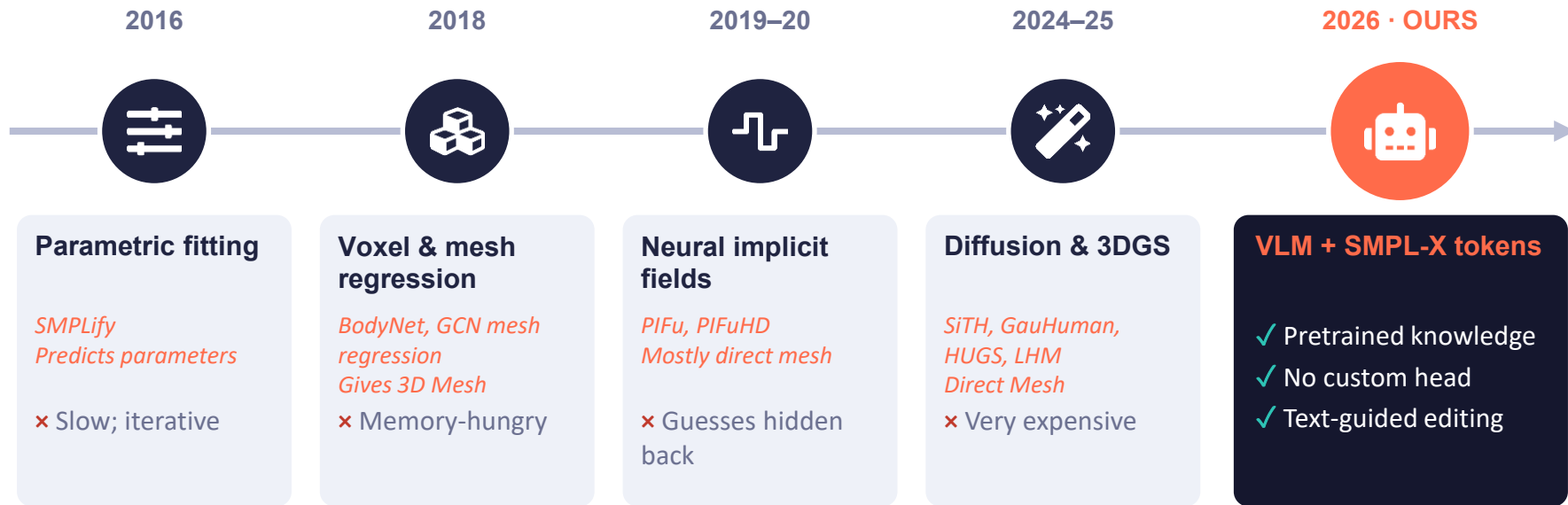
3

Translation

Position in 3D

122 numbers → **mesh with 10,475 vertices**

How the field has evolved



Our contribution: reconstruction as next-token generation — a pretrained VLM predicts SMPL-X tokens.

Contributions

Seq-to-seq pipeline



VLM fine-tuning

Qwen2.5-VL & InternVL2-8B



Visibility filtering

Frames $\geq 50\%$ visible



Tokenization

1,000 + 18 new tokens



Low-memory training

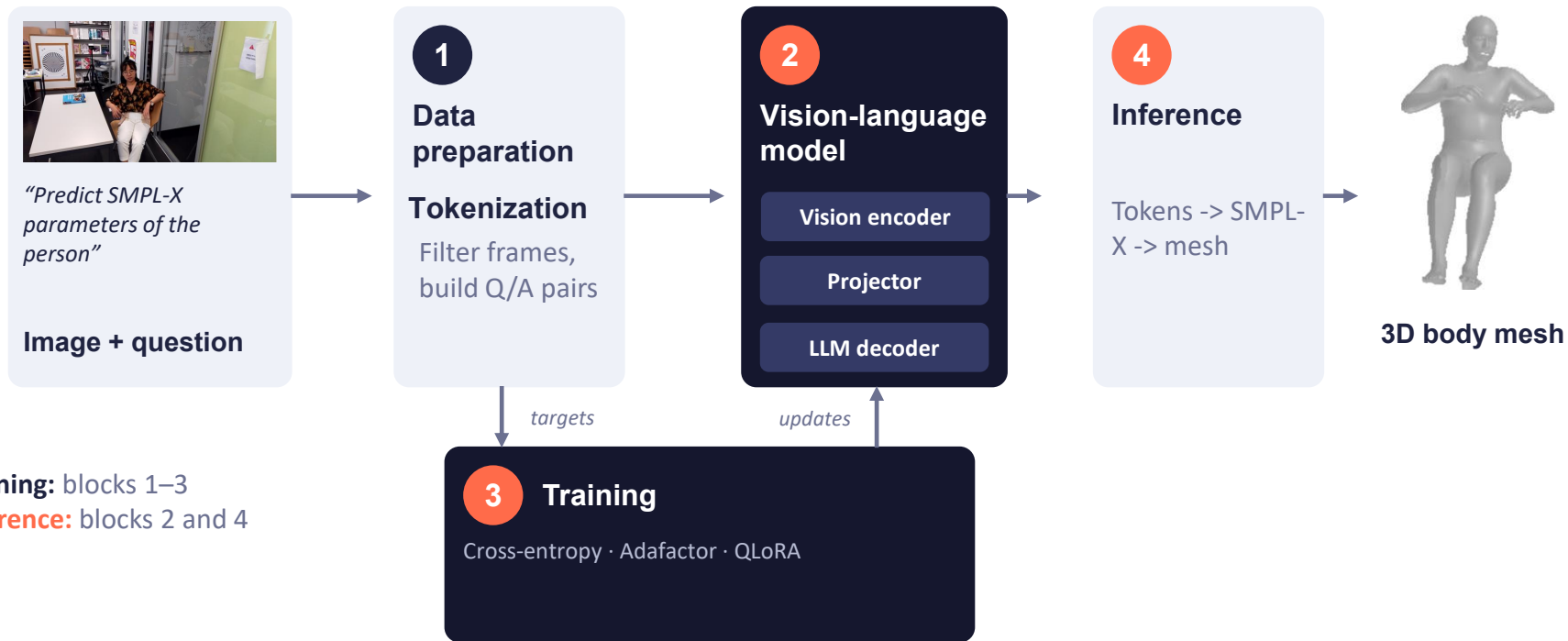
One 48 GB GPU



Competitive results

vs specialised baselines

System overview



Visibility-aware data filtering



Validity tag

Drop blurred / empty



Project body

2D Joint vertices →
Image



Visibility score

Visible ÷ total



Keep ≥ 50%

Human clearly visible

All frames **≥50% visible** 26% Improved 3D mesh
(MPJPE)

Image-question-answer pairs

1

Filter frame

≥50% visible

2

Load GT

EgoBody SMPL-X

3

Tokenize

122 → tokens

4

Add question

Natural language



→ 256 visual tokens

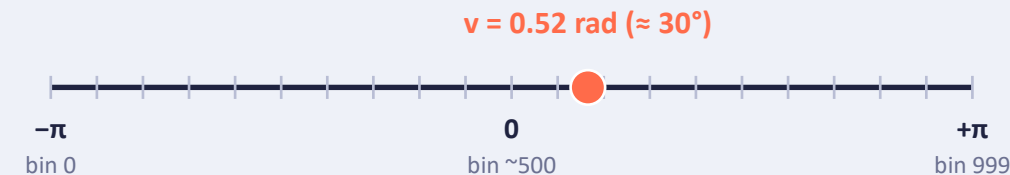
One training sample

```
{  
  "image": "frame_XXXXX.jpg",  
  "question": "<image> Predict SMPLX parameters of the person",  
  "answer": "<betas><pos_416>...</betas><body_pose>..."  
}
```

150,984 pairs · train 77,588 · val 20,280 · test 53,116

SMPL-X tokenization

A One value → one token



$$b = \text{round}((0.52 + \pi) / 2\pi \times 999) = 582$$

→ **<pos_582>**

Back: 582 → 0.519 rad ($\leq 0.18^\circ$ error)

B Range per group

Rotations $-\pi \dots \pi$

Shape, translation $-5 \dots 5$

Expression $-3 \dots 3$

C 122 values → one token sequence

1,000 positional + 18 structural = 1,018 new tokens



Group tag

Value token (one bin)

Two VLM backbones

InternVL2-8B

Full fine-tuning

Vision

InternViT-300M

Projector

Pixel-unshuffle + MLP → 256 tokens

Decoder

InternLM2.5-7B

Qwen2.5-VL-7B

QLoRA

Vision

Qwen2.5-VL ViT

Projector

2×2 merge + MLP → 256 tokens

Decoder

Qwen2.5-7B

Training the model on one 48 GB GPU



AdaFactor

~75% less optimizer memory



QLoRA

Small adapters, frozen 4-bit weights



Partial ViT tuning

Last 4 vision blocks + projector



Token-aware learning rate

Faster learning for new tokens

A40

48 GB GPU

12

batch size

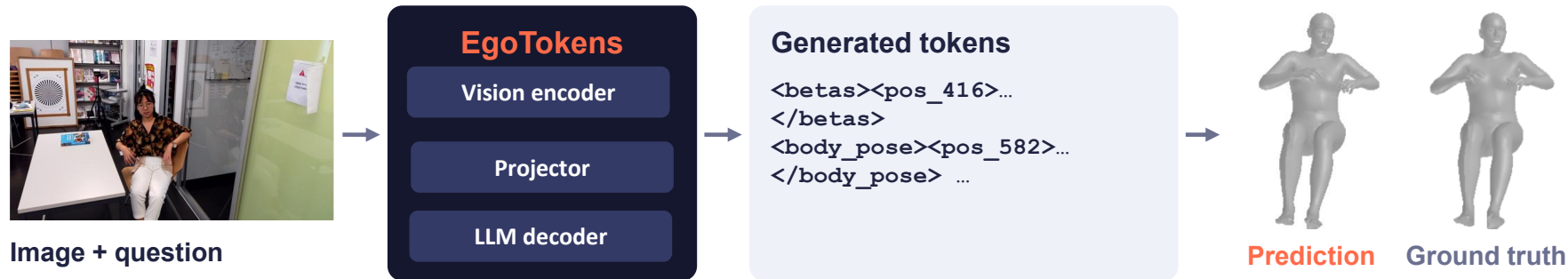
~100K

iterations

~250 ms

per image

Inference: from tokens to a 3D mesh



1 Read tokens

Tags mark each group

2 Dequantize

`<pos_582>` → 0.519 rad

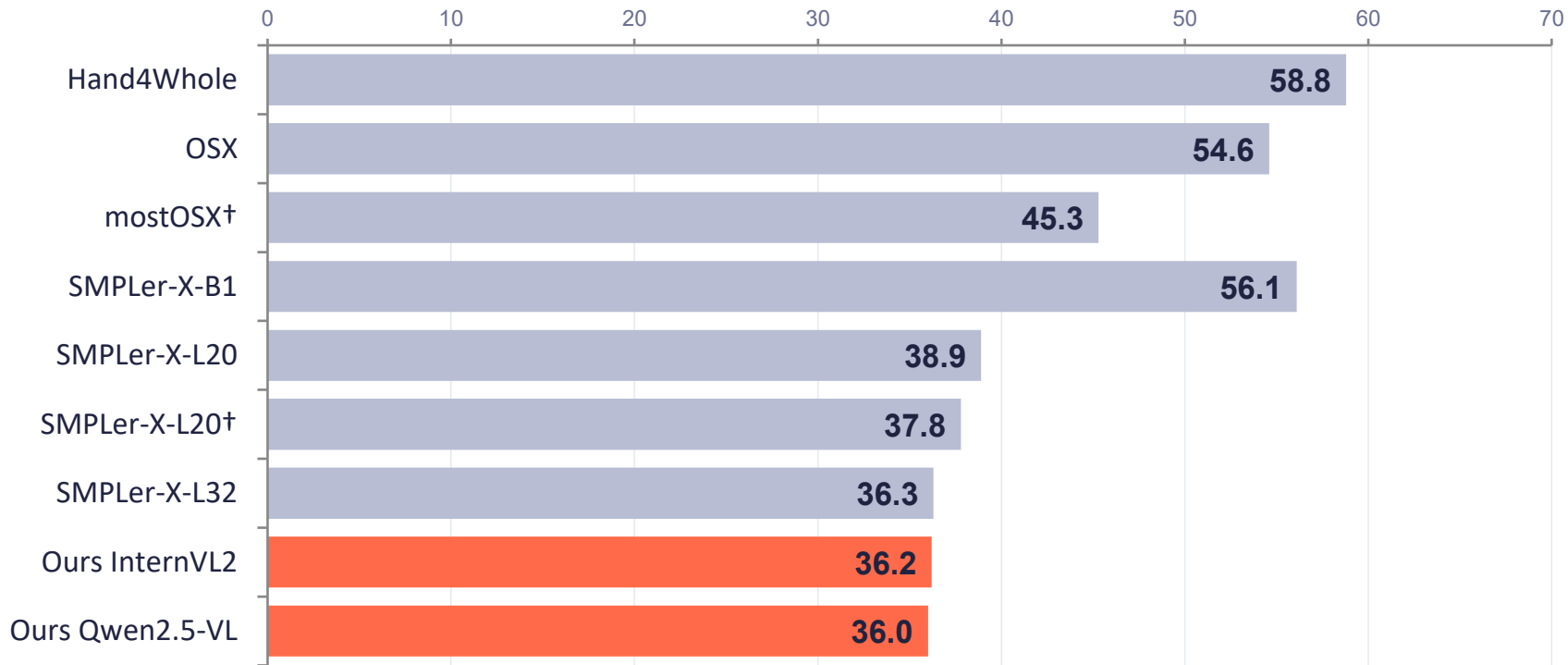
3 SMPL-X forward pass

122 values → 10,475-vertex mesh

≈ 250 ms per image · meshes from extended model

Results on EgoBody

PA-PVE (mm), lower is better

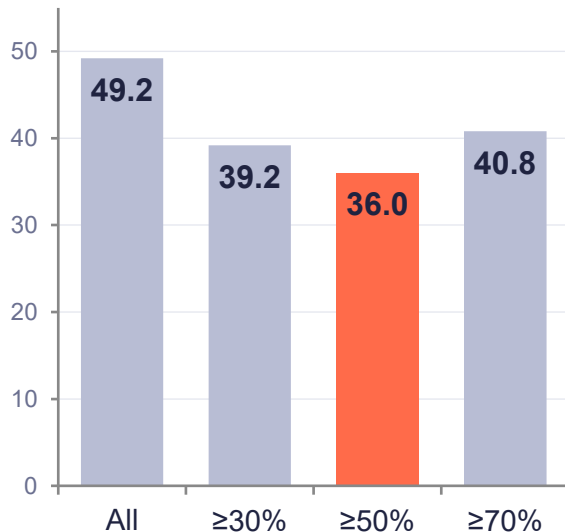


Ablation study

PA-PVE (mm), lower is better · coral = ours

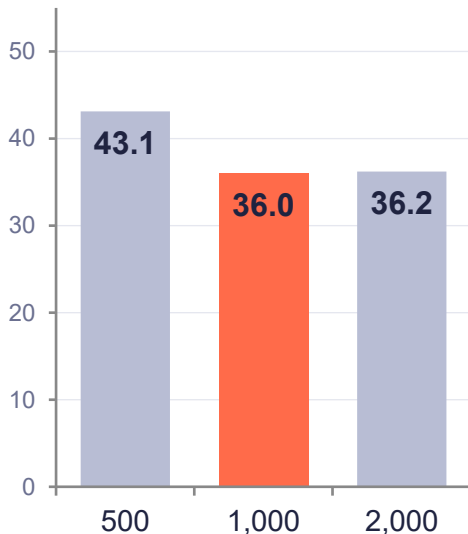
Visibility filtering

≥50% visible cuts error by 13 mm



Number of bins

1,000 is the sweet spot



Training choices

Loss

36.0 Cross-entropy

39.1 Focal

Learning rate

36.0 Separate for new tokens

37.3 Single

Optimizer

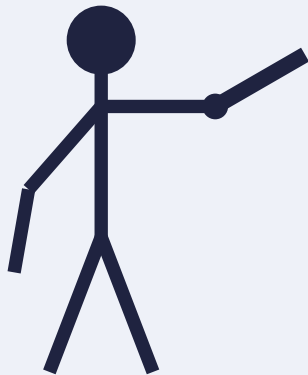
36.0 Adafactor

36.2 8-bit Adam

Ongoing 1: cross-entropy ignores distance

Problem

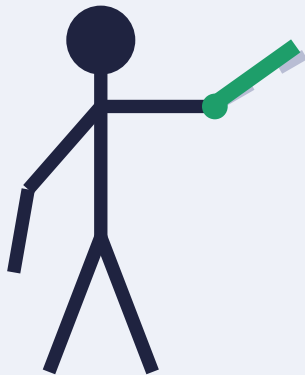
Ground truth



<pos_582> elbow 30°

Dashed = true arm. Both predictions give the true bin 10%.

Close prediction

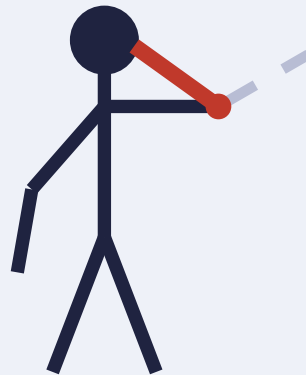


<pos_590> elbow 33°

Cross-entropy  2.30

Distance-aware L2  0.002

Far prediction



<pos_900> elbow 144°

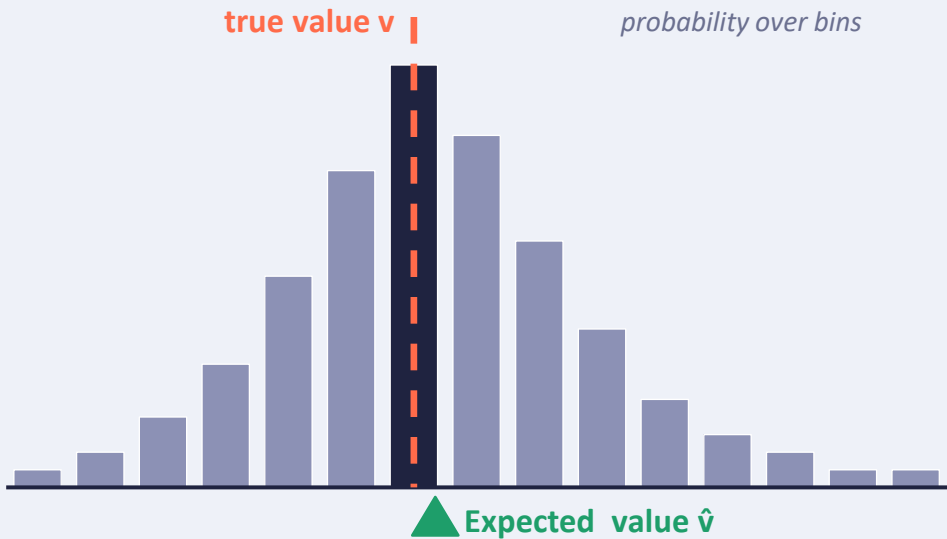
Cross-entropy  2.30

Distance-aware L2  3.24

Cross-entropy: same loss (2.30) for both mistakes. **Distance-aware L2:** the far pose costs over 1,000× more.

Ongoing 1: distance-aware loss

Solution



1

Softmax

over the 1,000 bins ($\tau = 0.8$)

2

Expected value

$$\hat{v} = \sum p_k \cdot b_k$$

3

L2 loss

$(\hat{v} - v)^2$ grows with distance

$$\mathcal{L} = 0.3 \cdot \text{CE} + 0.7 \cdot \text{L2}$$

CE keeps tags correct

Ongoing 2: language-guided shape editing

Problem: the model can only reconstruct · Solution: one model that also edits from text



Input image

"...with a fat belly" / "...with a shorter height"



Predicted



Belly +



Belly -



Height +



Height -

1 Select region

Belly, legs, face or height

2 Scale vertices

Grow or shrink the region

3 Fit $\Delta\beta$

Change in shape coefficients

8

edits per image

150K → 1.36M

training samples

Published EgoTokens (An extension of this work)

- Published in Machine Learning with Application
Journal IF 6.1

<https://doi.org/10.1016/j.mlwa.2026.101042>



Machine Learning with Applications
Volume 26, December 2026, 101042



Recommended articles
No articles found.

EgoTokens: Egocentric 3D human reconstruction and language-guided shape editing via quantized position tokens in vision-language models

Gulraiz Khan , Maneeha Rani , Waqas Ahmed , Kevin Pimblett , Kenneth Y. Wertheim 

Show more 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.mlwa.2026.101042>  [Get rights and content](#)  [View accessibility information](#) 

Under a Creative Commons license   Open access 

Abstract

3D human body reconstruction from egocentric images is a challenging task because of environmental and self-occlusion. Downward viewpoints and scale ambiguity can also hinder precise 3D estimation. Current methods mainly rely on specialized regression networks to reconstruct 3D meshes that often lack structural guidance from high-

Takeaways

Pretrained VLMs can reconstruct 3D humans by generating quantized SMPL-X tokens.
End-to-End model can understand and edit the 3D model without any additional head.

Next: distance-aware loss and text-driven shape editing in the same model.

Thank you — questions?

G.Khan-2022@hull.ac.uk