

FlowAvatar: Real-Time Full-Body Avatars from Sparse Egocentric Inputs on Consumer XR Devices

Chenfeng Gao^{1*}, Taeyoung Yeon¹, Sungheon Park¹, Vasco Xu^{2†}, Anish Prabhu^{3‡},
Mar Gonzalez-Franco³, Karan Ahuja¹

¹Northwestern University ²University of Chicago ³Google



Figure 1: FlowAvatar: a unified system for real-time full-body pose tracking, hand and facial expression tracking, and personalized look-alike appearance on consumer XR devices.

ABSTRACT

Prior avatar embodiment research has studied tracking completeness (how faithfully the body, hands, and face are sensed and reproduced) and personalized appearance in isolation, each requiring specialized lab equipment. How these factors interact on consumer hardware remains an open question, because no existing system unifies body pose, face, hands, and personalized appearance on a standalone headset. We present FlowAvatar, a system enabling real-time personalized look-alike avatars with full-body pose, hand articulation, facial expression, and eye gaze tracking from sparse egocentric inputs on consumer XR devices (Meta Quest Pro). Our pose model achieves competitive spatial accuracy and state-of-the-art temporal consistency on the AMASS benchmark against prior works with comprehensive spatial and temporal losses validated through architecture and loss ablations, while running in real-time. Two user studies (N=12, N=10) provide converging evidence that tracking completeness and personalized appearance serve complementary roles: tracking completeness broadly improves perceived motion quality, while only their combination achieves positive total embodiment.

Keywords: User Digitization, Full-Body Pose, XR Avatars

1 INTRODUCTION

In virtual environments, people act through an avatar [58], and how strongly they feel that avatar as their own body governs much of

the experience. This sense of embodiment rests on three components: body ownership, agency, and self-location [32]. When it is strong, users feel more present and extended reality (XR) becomes more effective for social interaction, training, and productivity [55]. Unlike immersion, which is the objective sensorimotor fidelity the hardware affords [57], embodiment is a response of the user. A growing body of work points to which bodily cues shape it: given a free choice, users building their ideal avatar prioritize faithful control over appearance [20]; adding tracked points such as the feet and pelvis to a head-and-hands avatar raises embodiment [18]; and even simple lip-sync of the face strengthens self-identification [24]. Together these results point to *tracking completeness*, how many body segments are sensed and faithfully reproduced, as a central driver of embodiment, alongside personalized appearance [22] and the temporal coherence of motion [2, 12].

However, each of these studies isolated a single factor and required instrumentation beyond a standalone headset [18, 20, 24]. Two questions remain open. First, how do tracking completeness and personalized appearance each shape embodiment when *combined* in one avatar? Prior work leans toward control over appearance but never tested face and hand completeness alongside personalized texture [20]. Second, do these findings hold when tracking relies on a headset’s own sensing, which is sparser and noisier and where temporal artifacts such as jitter especially undermine embodiment [23]?

Because these two factors interact rather than act independently, they cannot be teased apart by mixing separate modular systems: swapping components changes several variables at once. Only a single pipeline can hold one factor *fixed* while varying the other. Yet no consumer system unifies full-body tracking, hands, face, and personalized appearance with sufficient temporal quality (Table 1): Meta lacks face tracking [43], Apple lacks full-body pose [6], and research systems optimize isolated subsystems.

We present FlowAvatar, a unified system that enables real-time personalized look-alike avatars with full-body pose, hands, facial expressions, and eye gaze tracking on off-the-shelf consumer XR devices (Figure 1). A one-time enrollment generates personalized

*e-mail: chenfenggao2029@u.northwestern.edu, {taeyoungyeon, sungheon.park, kahuja}@northwestern.edu. C. Gao and T. Yeon contributed equally to this work.

†e-mail: vascoxu@uchicago.edu

‡e-mail: {anishprabhu, margon}@google.com

Table 1: Comparison with related work for user digitization on consumer XR devices without additional instrumented sensors.

Method	Body Pose	Face	Hands	Personalized	On-Device	MPJPE	MPJVE	Jitter
Meta Quest Full Body [43]	✓	×	✓	×	✓	—	—	—
Apple Vision Pro Persona [6]	×	✓	✓	✓	✓	—	—	—
Pixel Codec Avatar [39]	×	✓	×	✓	✓	—	—	—
AvatarPoser [30]	✓	×	×	×	×	6.62	33.16	10.79
AGRoL [17]	✓	×	×	×	×	6.74	24.14	6.33
AvatarJLM [72]	✓	×	×	×	×	5.96	27.50	6.91
HMDPoser [14]	✓	×	×	×	✓	5.44	30.15	5.62
FlowAvatar (our work)	✓	✓	✓	✓	✓	5.40	21.72	3.77

textures that capture the user’s appearance from a brief smartphone video. At runtime, body pose from our neural network, articulated hands and facial expressions from Meta Quest Pro’s built-in tracking, and eye gaze all drive a single SMPL-X [49] mesh simultaneously without additional sensors. Our pose model combines body-part encoding, unified kinematic decoding, and comprehensive spatial and temporal losses, achieving competitive spatial accuracy with state-of-the-art temporal consistency while running in real time.

Two user studies provide converging evidence on the roles of tracking and appearance in embodiment. In a motion quality study (N=12), adding face and hand tracking significantly improved all six perceived quality dimensions, while adding personalized texture reached significance only for Interaction Readiness. In an embodiment study (N=10), tracking completeness was associated with *agency*, while personalized appearance was associated with *ownership*, and only the condition combining both achieved positive total embodiment. These results suggest that tracking completeness and personalized appearance serve complementary roles, and that platform designers aiming for full embodiment should invest in both rather than treating them as competing priorities.

This paper contributes:

1. **FlowAvatar**, the first open-source avatar system to holistically integrate real-time full-body pose, hand articulation, facial expression, and eye gaze tracking with personalized appearance on a standalone consumer XR device, enabling the study and evaluation of a complete avatar under realistic consumer constraints.
2. A **full-body pose framework** that infers pose from sparse egocentric head and hand inputs through body-part encoding, unified kinematic decoding, and comprehensive spatial and temporal losses, achieving competitive spatial accuracy and state-of-the-art temporal consistency on the AMASS dataset, and released as open source.
3. **Two user studies**, enabled by this integrated pipeline, that disentangle the complementary roles of tracking completeness and personalized appearance in shaping perceived motion quality and embodiment.

2 RELATED WORK

Our work sits at the intersection of graphics, perception, and extended reality body capture. We are not the first to recognize the value of digitizing user bodies, capturing personal appearance, and creating embodied avatars. Thus, we first succinctly review related work in motion capture, sensed both externally and by worn hardware. We then discuss research efforts focused towards user digitization from XR headsets, and their effects on embodiment and avatar representation.

2.1 User Digitization from External Sensors

There has been significant progress in generating digital representations using external sensors and capture rigs. Commercial mo-

tion capture systems such as Vicon use specialized hardware, worn retroreflective markers, and multi-camera rigs to create millimeter-accurate digitizations. Methods such as Holoportation [48] and Starline [36] employ a suite of external sensors (RGB, depth, IR) to generate photorealistic representations of humans and environment. While compelling, their prohibitive costs make them impractical for everyday use.

Therefore, much research has been devoted to instrumentation-free approaches that rely on computer vision techniques. Numerous works have looked into particular verticals such as tracking eyes, hands, or the face [7]. For unified full-body tracking, works have predominantly made use of a combination of depth and RGB [21] cameras.

Photorealistic avatar creation from RGB has progressed from mesh-based texture reconstruction [73] to neural rendering methods that generate animatable avatars from monocular video [26, 31], with multi-camera setups further enhancing quality and drivability [52]. Recent work has also begun distilling these high-fidelity Gaussian avatars into mobile-ready representations, enabling real-time rendering of multiple avatars on standalone VR headsets [27].

2.2 User Digitization from Body Worn Sensors

Wearable sensors provide a portable alternative to external sensors. Since wearables can be intrusive, research has often focused on digitizing specific body parts. Hands, for example, have been digitized through wrist-worn cameras [67], EMG [38], EIT [34], antennas [33], and depth armbands [16]. For full-body pose, body-worn IMUs are most common. Commercial systems such as Xsens use many high-grade inertial sensors (typically 17), and later work has reduced this count to 6 [69] and even 3 [63]. More recently, work has explored using IMUs found in everyday consumer devices [45, 68], albeit producing a lower-fidelity full-body pose estimate.

2.3 User Digitization from XR Headsets

Many works instrument either the headset [1, 10], or the wearer [14] with additional sensors to digitize parts of their body in XR. Although these approaches can achieve high fidelity, they typically require significant server-side GPU compute, increase system complexity, and reduce user comfort. This has motivated a shift towards instrumentation-free user digitization using only the sensing capabilities of standard headsets.

Recently, instrumentation-free methods have advanced specific digitization components: photorealistic face capture [52] and mesh-based body reconstruction [3]. The works most related to ours estimate full-body pose from an XR headset’s sparse observations (head + hands). Approaches such as AvatarPoser [30], HMDPoser [14], and AGRoL [17] use data-driven models with SMPL [8], often assuming continuous hand visibility in the headset’s FoV. To relax this assumption, EgoPoser [29] incorporates random hand masking or FoV modeling, but typically relies on spatio-temporal transformers with substantial computational cost, which complicates real-time on-device deployment.

Across these methods, a trend toward increasing architectural complexity is evident: from RNNs [30] to spatio-temporal Transformers [29] to conditional diffusion models [17]. Besides architecture choice, these methods also vary in their use of temporal supervision: HMDPoser [14] employs an acceleration loss, AvatarJLM [72] uses multi-scale velocity losses at multiple frame strides, while AvatarPoser [30], AGRoL [17], and EgoPoser [29] rely solely on spatial objectives. Our work combines both acceleration and velocity supervision with a lightweight recurrent architecture, and we ablate both dimensions independently in Section 5.1.2. FlowAvatar builds on this body of work as a unified system connecting personalized appearance generation to real-time body, hand, face, and eye tracking on consumer XR hardware.

2.4 Embodiment and Avatar Representation

Avatars are the primary channel for self-representation in XR. Higher-fidelity avatars increase trust and engagement from third-person perspectives [28] and strengthen first-person embodiment [50], which has been linked to improved empathy [51], presence [56], and cognitive performance [59]. The single-factor studies noted earlier [18, 20, 24] exemplify a broader literature that a recent meta-analysis of body ownership illusions synthesizes: both tracking fidelity and avatar appearance reliably shape embodiment, yet their relative contributions remain hard to compare because studies manipulate them in isolation and under heterogeneous conditions [46].

Beyond these findings, Debarba et al. [15] showed that hands and face contribute disproportionately to perceived animation plausibility compared to gross body segments. On the appearance side, look-alike avatars elicit greater neural self-recognition than generic ones [22], with personalized avatars increasing self-identification over cartoon representations [53]. Tao et al. [61] found physics-aware avatar motion can improve embodiment even with small motion discrepancies. Even recent work examines these factors one at a time: Tschanter et al. [62] find that the value of embodied tracking depends on the task, while Fiedler et al. [19] show that avatar appearance and its alignment with the user’s real body influence self-perception and behavior. Personalized look-alike avatars are meanwhile becoming widely accessible through smartphone-based reconstruction, removing the need for capture rigs [42].

FlowAvatar brings these separately studied components into one driven avatar on consumer hardware, letting us study tracking completeness and personalized appearance together rather than one at a time, a comparison prior work could not make (Section 6).

3 POSE ESTIMATION PIPELINE

FlowAvatar builds and drives a personalized avatar in two stages: a one-time enrollment that generates the look-alike avatar from a short phone video, and a real-time runtime that animates it on the headset. We present the runtime pose estimation first, as the technical core, and cover enrollment and deployment in Section 4. The avatar is represented using the SMPL-X parametric body model [49], a single mesh controlled by body shape (β) for proportions, body pose (θ) for joint rotations, hand articulation, and facial blendshapes for expressions. This section describes how we animate that mesh in real time from sparse headset signals: a neural network infers full-body pose from head and hand tracking (Section 3.1.1–Section 3.1.3), and the estimated body pose is then combined with hand, face, and eye tracking from the headset’s built-in sensors (Section 3.2).

3.1 Body Pose Estimation

3.1.1 Input and Output Representation

Following prior work [14, 30], we represent each tracked component with position ($p \in \mathbb{R}^3$) in the global coordinate system, orientation ($r \in \mathbb{R}^6$) using the 6D rotation representation, angular velocity ($\omega \in \mathbb{R}^6$) and linear velocity ($v \in \mathbb{R}^3$) computed via finite differences.

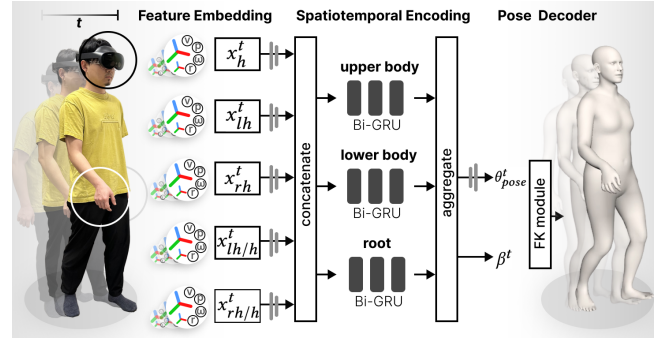


Figure 2: Pose estimation architecture: three parallel BiGRU networks (upper body, lower body, root) process shared embedded features to generate SMPL-X parameters.

For each frame t , we construct input vectors for the head ($x_h^t \in \mathbb{R}^{18}$), left hand ($x_{lh}^t \in \mathbb{R}^{18}$), and right hand ($x_{rh}^t \in \mathbb{R}^{18}$), where each vector concatenates position, orientation, angular velocity, and linear velocity. Additionally, we compute head-relative representations of the hand positions ($x_{lh/h}^t, x_{rh/h}^t \in \mathbb{R}^{18}$), which provide valuable cues for estimating body proportions and improving joint position accuracy. The complete input at time t is:

$$x^t = [x_h^t, x_{lh}^t, x_{rh}^t, x_{lh/h}^t, x_{rh/h}^t] \in \mathbb{R}^{90} \quad (1)$$

To capture temporal dynamics, we process sequences of $T = 40$ consecutive frames during both training and inference, corresponding to ~ 667 ms at our device capture frequency of 60 Hz. During inference, we employ a sliding window approach where each prediction uses the current frame and the previous $T - 1$ frames, enabling real-time performance while maintaining temporal consistency. The system outputs SMPL-X [49] body parameters including joint rotations ($\theta_{pose}^t \in \mathbb{R}^{6 \times 22}$) and shape parameters ($\beta \in \mathbb{R}^{10}$), which are processed through forward kinematics to drive the final 3D mesh with articulated pose.

3.1.2 Architecture

For inputs $x^t \in \mathbb{R}^{90}$ at time t , we first apply a feature embedding module to transform these sparse tracking signals into higher-dimensional representations $f_e^t = \phi_e(x^t) \in \mathbb{R}^{d_e}$, where ϕ_e is implemented as a two-layer MLP with ReLU activations and $d_e = 256$. We concatenate the raw inputs before embedding for parameter efficiency.

After feature embedding, we process the embedded features through three parallel bidirectional GRU networks, each specialized for a different body region (upper body, lower body, and root). All three GRUs receive the same embedded input features but learn region-specific temporal dynamics:

$$h_t^{part} = \text{BiGRU}_{part}(f_e^t, h_{t-1}^{part}) \quad (2)$$

where each h_t^{part} represents the hidden state for a specific body region at time t and $d_h = 512$. Together, all three GRUs model the spatio-temporal patterns across all body regions. The outputs are then fed into a unified decoder to generate the SMPL-X body parameters:

$$\theta_{pose}^t = \phi_{pose}(h_t) \quad (3)$$

The decoder ϕ_{pose} consists of a two-layer MLP with ReLU activations and dropout ($p = 0.1$) that outputs the complete pose parameters $\theta_{pose}^t \in \mathbb{R}^{6 \times 22}$ for all body joints at time t . Along with the predicted shape parameters $\beta^t \in \mathbb{R}^{10}$, these are processed through SMPL-X forward kinematics to generate the final 3D joint positions

$P^t \in \mathbb{R}^{22 \times 3}$ and the corresponding mesh with articulated pose. We select GRU over Transformer as it achieves better accuracy across all metrics under identical training configurations (Section 5.1.2).

3.1.3 Training Objectives

We employ multiple complementary loss functions organized into four categories to guide our model toward accurate, temporally consistent motion.

Pose loss. We employ L1 root orientation loss ($\mathcal{L}_{ori}$), L1 local joint rotation loss ($\mathcal{L}_{lrot}$), L1 global joint rotation loss ($\mathcal{L}_{grot}$), and L1 joint position loss ($\mathcal{L}_{pos}$), following previous approaches [14, 29]. The local rotation loss ensures accurate joint angles, while the global rotation loss helps reduce error accumulation along the kinematic chain. Joint position loss provides direct supervision on the spatial configuration of the body.

Temporal loss. For temporal consistency, we adopt an acceleration loss that penalizes deviations in joint accelerations between predicted and ground-truth motions:

$$\mathcal{L}_{acc} = \frac{1}{(T-2) \cdot J} \sum_{t=1}^{T-2} \sum_{j=1}^J \|a_j^t - \hat{a}_j^t\|_1 \quad (4)$$

where a_j^t is the acceleration of joint j at time t , computed via second-order finite differences. Additionally, we implement a multi-scale velocity loss that measures velocity differences at varying time scales:

$$\mathcal{L}_v(l) = \frac{1}{(T-l) \cdot J} \sum_{t=1}^{T-l} \sum_{j=1}^J \| (P_j^{t+l} - P_j^t) - (\hat{P}_j^{t+l} - \hat{P}_j^t) \|_1 \quad (5)$$

Hand alignment loss. Standard inverse kinematics post-processing [30] iteratively adjusts predicted joint angles to match observed hand positions at inference time, adding computational overhead and introducing discontinuities between corrected and uncorrected frames. We instead integrate a learned hand alignment loss that maintains end-to-end trainability:

$$\mathcal{L}_h = \frac{1}{2T} \sum_{t=1}^T (\|P_{lh}^t - \hat{P}_{lh}^t\|_1 + \|P_{rh}^t - \hat{P}_{rh}^t\|_1) \quad (6)$$

where P_{lh}^t and P_{rh}^t are the predicted positions of left and right hands, ensuring accurate hand placement without runtime post-processing.

Ground contact loss. We employ two complementary losses for ground contact regularization. The foot contact loss enforces zero velocity when feet are in contact with the ground:

$$\mathcal{L}_{fc} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|(P_{feet}^{t+1} - P_{feet}^t) \cdot m^t\|_1 \quad (7)$$

where $m^t \in \{0, 1\}^{n_{feet}}$ is a binary foot contact mask, preventing the common ‘‘skating’’ artifact. We also implement a ground penetration loss:

$$\mathcal{L}_p = \frac{1}{T} \sum_{t=1}^T \|(z_{min}^t - z_{ground}) \cdot l^t\|_1 \quad (8)$$

where z_{ground} is the ground height, z_{min}^t is the height of the lowest joint, and $l^t \in \{0, 1\}^J$ indicates whether a joint is below ground. Together, these constraints address the two most common artifacts in egocentric pose estimation, foot skating and ground penetration, which we evaluate quantitatively in Section 5.1.2.

Our complete loss function is defined as:

$$\begin{aligned} \mathcal{L} = & \lambda_{ori} \mathcal{L}_{ori} + \lambda_{lrot} \mathcal{L}_{lrot} + \lambda_{grot} \mathcal{L}_{grot} + \lambda_{pos} \mathcal{L}_{pos} + \\ & \lambda_{acc} \mathcal{L}_{acc} + \lambda_{vel} \mathcal{L}_{vel} + \lambda_h \mathcal{L}_h + \\ & \lambda_{fc} \mathcal{L}_{fc} + \lambda_p \mathcal{L}_p \end{aligned} \quad (9)$$

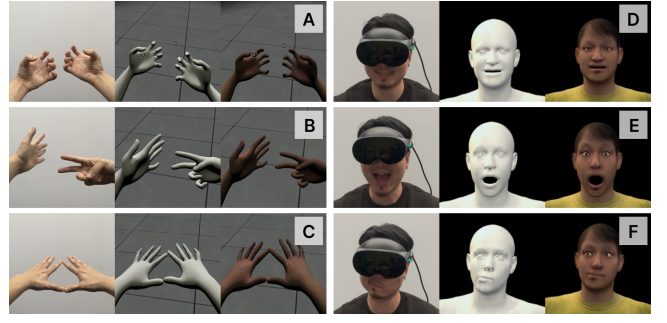


Figure 3: Examples of hand (A–C) and facial expression (D–F) capture on Meta Quest, remapped to SMPL-X mesh.

3.2 Hand, Face, and Eye Pose Integration

Beyond body pose, the Quest Pro’s built-in sensors provide hand articulation, facial expressions, and eye gaze. Each of these operates in a different parameter space from SMPL-X, so integrating them onto the same mesh requires mapping for each modality (Figure 3).

For hand tracking, we perform a one-time offline rig setup in Autodesk Maya, retargeting Meta’s 18-joint OVR hand skeleton (with an extra joint on the thumb and pinky) onto SMPL-X’s 16-joint hand (15 finger joints plus wrist) to produce a rigged SMPL-X mesh that we then import into our Unity runtime. At runtime, the retargeted hand poses blend with the body pose from our neural network, producing coordinated full-body motion with articulated fingers.

For facial expressions, at runtime we map Meta’s 52 ARKit-compatible blendshape coefficients [5] to SMPL-X’s expression space using a conversion matrix from EMAGE [37], giving 100 expression parameters plus 3 jaw rotations. Because our Unity runtime exposes only 10 expression blendshapes, we reduce these 100 coefficients to those 10 through an empirically calibrated projection, and pass the jaw rotations through directly. This trades subtle detail for real-time rendering while preserving the most salient movements such as smiling, frowning, mouth opening, and brow movement.

Additionally, we extract eye movement data from Meta’s gaze tracking to control SMPL-X eyeball rotations (yaw and pitch for each eye), enabling natural eye contact in social scenarios.

4 FULL-BODY AVATAR PIPELINE

With the pose estimation pipeline in place, we now describe how FlowAvatar produces and delivers the complete look-alike avatar in two stages: a one-time enrollment that builds a personalized avatar offline from a brief smartphone video (Section 4.1), and a runtime that drives it on the headset (Section 4.2). We build on established differentiable rendering [60] and enroll each user from a single phone video, the standard, low-friction workflow that commercial systems such as Meta Horizon [44] use to keep avatar creation accessible.

4.1 Enrollment and Texture Generation

The user opens a companion smartphone app and records a ~ 10 -second video while slowly completing a full turn using the front-facing camera. No special lighting, background, or calibration is required. From this video, our pipeline generates a personalized UV texture map and estimates initial body-shape (β) parameters through the following four stages.

A. Preprocessing. We estimate SMPL-X shape, pose, and camera parameters for each frame (~ 250 frames at 25 FPS) using PyMAF-X [70], and segment body regions from 2D pose [13] and a human segmentation mask [11]. To mitigate inaccuracies in camera parameter estimation, we adopt a three-component texture design, training separate textures for the body, face, and head using corresponding

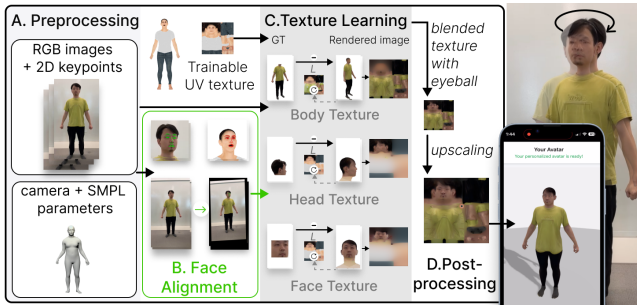


Figure 4: Texture generation pipeline: (A) preprocessing, (B) face alignment, (C) region-specific differentiable rendering, (D) compositing and upscaling.

region masks, so each region can be aligned and optimized independently.

B. Face alignment. PyMAF-X assumes a weak-perspective camera model, which causes misalignment between rendered and input images, particularly degrading face texture quality under perspective distortion. We correct this with a 2D affine transform computed from facial landmarks [74] detected in both rendered and input images, aligning the input to serve as ground truth for face texture training.

C. Texture learning. Using the estimated SMPL-X and camera parameters, we render 2D images with trainable UV textures via a differentiable renderer [35] and optimize against the input RGB frames (Figure 4). Following Svitov et al. [60], the texture loss combines L_2 reconstruction, LPIPS [71], SSIM, and total variation regularization. Textures for the face, head, and body are trained separately with corresponding masks, typically converging within 500 iterations.

D. Postprocessing. The learned face, head, and body textures are blended into a single 512×512 texture, with an eyeball texture from the reference template composited into the final map. The result is upscaled to 2048×2048 via Real-ESRGAN [65] to reduce noise and improve detail. Shape betas from the texture generation process are applied to the SMPL-X mesh at runtime to customize body proportions. Total processing time is approximately 15 minutes on an NVIDIA RTX 4090. The user can preview the resulting avatar on the phone before proceeding to XR.

4.2 Runtime and Deployment

The user puts on a Quest Pro and logs in to fetch the avatar texture generated during enrollment (Section 4.1). No calibration or T-pose is needed. At each frame, the pose estimation model infers full-body pose from the headset’s head and hand tracking (Section 3), while the Quest Pro’s built-in sensors provide hand articulation, facial expressions, and eye gaze. These are combined onto the same SMPL-X mesh (Section 3.2), producing a coherent avatar with synchronized body, hands, face, and eyes (Figure 1). The model also predicts shape (β) parameters, which are used once at startup to refine the body proportions estimated during enrollment.

FlowAvatar supports two deployment configurations, both of which we release as open source at <https://github.com/SPICeXLAB/FlowAvatar>: a streaming configuration and an on-device configuration.

In the **streaming configuration**, Meta Quest Pro tracking data is streamed to a Windows laptop with an RTX 4090 for inference. The end-to-end pipeline runs at 247 FPS under this setup; HMDPoser [14] achieves 43.3 FPS under the same configuration. The pose model alone runs at 91.5 FPS on a general-purpose CPU (Apple M1 Max), increasing to 341.0 FPS after ONNX optimization ($3.7\times$). All user studies reported in this paper use the streaming configuration.

In the **on-device configuration**, FlowAvatar runs fully on Meta Quest Pro (CPU) via Unity Sentis’s ONNX runtime. Because Unity Sentis lacks GRU operator support, the on-device model uses the LSTM variation, a modest accuracy drop (MPJRE 4.43 vs. the GRU’s 4.19; Table 4). We sample head, hand, and face into a circular buffer and run capture and inference asynchronously. The complete pipeline sustains 27 FPS at a 97 ms end-to-end motion delay, enabling practical untethered use. The hand (~ 50 ms) and face (~ 70 ms) trackers run asynchronously on native Meta threads and the pose loop reads their latest state, so the slowest tracker sets the delay but not the frame rate. This motion delay stays well within the ~ 130 – 170 ms that prior work finds tolerable for avatar motion [25]. We have also deployed the GRU model on AndroidXR (Samsung Moohan, Snapdragon XR2 Gen 2) via TFLite at 60 Hz with no accuracy loss.

5 EVALUATION

Before examining how tracking completeness and personalized appearance affect embodiment (Section 6), this section validates the pose estimation model through benchmark comparison, ablation studies, and input robustness (Section 5.1), followed by a preliminary texture generation evaluation (Section 5.2).

5.1 Full-body Pose Estimation

Following recent common practice [14, 17, 30, 72], we evaluate on the public AMASS dataset [40] using the train/test split introduced by HMDPoser [14]: twelve AMASS subsets for training with HumanEva [54] and Transition held out for testing, and ground-truth body shape parameters used to compute joint positions. This split tests generalization to entirely unseen motion datasets rather than random partitions within the same subsets. We train with Adam optimizer (initial learning rate 10^{-4} decaying to 10^{-5} after 75% of training), batch size 256, for 400 epochs on an RTX 4090. For our loss (Equation (9)), we set $\lambda_{ori} = 1.0$, $\lambda_{rot} = 5.0$, $\lambda_{grot} = 1.0$, $\lambda_{pos} = 5.0$, $\lambda_{acc} = 0.5$, $\lambda_{vel} = 5.0$, $\lambda_h = 5.0$, $\lambda_{fc} = 1.0$, and $\lambda_p = 1.0$.

All benchmark results below use the streaming configuration described in Section 4.2 with the 3-layer BiGRU model. Following established practices [14, 17], we report spatial metrics, including MPJRE (mean joint rotation error in degrees) and MPJPE (mean joint position error in cm), alongside temporal metrics, including MPJVE (mean joint velocity error in cm/s) and Jitter (third derivative of position, 10^2 m/s³), which reflects the smoothness of generated motion.

5.1.1 Comparison with Prior Work

Table 2 shows that FlowAvatar achieves competitive spatial accuracy and state-of-the-art temporal consistency on the AMASS benchmark. Compared to HMDPoser [14], FlowAvatar matches its spatial accuracy (MPJPE: 5.40 cm vs 5.44 cm) while substantially improving temporal metrics: MPJVE of 21.72 cm/s and jitter of $3.77 \cdot 10^2$ m/s³ represent 28% and 33% improvements respectively over HMDPoser’s 30.15 cm/s and $5.62 \cdot 10^2$ m/s³. AvatarPoser [30] shows high jitter ($10.79 \cdot 10^2$ m/s³) and MPJVE (33.16 cm/s), and AvatarJLM [72] likewise has elevated jitter ($6.91 \cdot 10^2$ m/s³) and velocity

Table 2: FlowAvatar compared with prior works on AMASS.

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jit.↓
AvatarPoser [30]	4.68	6.62	33.16	10.79
AGRoL [17]	4.38	6.74	24.14	6.33
AvatarJLM [72]	4.45	5.96	27.50	6.91
HMDPoser [14]	4.27	5.44	30.15	5.62
EgoPoser [29]	4.45	6.85	35.20	18.64
FlowAvatar (ours)	4.19	5.40	21.72	3.77

error (27.50 cm/s). AGRoL [17] reaches the closest temporal metrics among the baselines but higher positional error (MPJPE: 6.74 cm). EgoPoser [29] shows the highest velocity error (35.20 cm/s) and jitter ($18.64 \times 10^2 \text{ m/s}^3$, nearly $5 \times$ ours), reflecting a focus on large-space global positioning over temporal smoothness. Since spatial offsets below 10 cm are typically imperceptible [12], all methods fall within the perceptual threshold on position, making temporal artifacts the primary differentiator.

5.1.2 Ablation Studies

Loss ablation. Table 3 shows the impact of each loss component. Removing temporal losses increases MPJVE from 21.72 to 25.58 cm/s and jitter from 3.77 to $5.17 \times 10^2 \text{ m/s}^3$, representing 18% and 37% degradation respectively, the largest single-category impact. Without hand alignment loss, H-PE increases from 2.31 cm to 2.98 cm (29% worse). The ground contact losses show trade-offs: their removal slightly improves hand accuracy (2.25 cm vs 2.31 cm) but degrades jitter (4.05 vs $3.77 \times 10^2 \text{ m/s}^3$) and leads to visually noticeable foot-floating artifacts during high-intensity motions. The combination of all losses achieves the best overall balance.

Table 3: Ablation across different loss configurations.

Strategy	MPJRE↓	MPJVE↓	Jitter↓	H-PE↓
No temporal loss	4.55	25.58	5.17	2.50
No hand loss	4.41	23.26	3.91	2.98
No ground loss	4.31	22.55	4.05	2.25
All losses	4.19	21.72	3.77	2.31

Backbone comparison. Table 4 compares four backbone architectures under identical training configurations, with FPS measured on an Apple M1 Max CPU for model-only inference. Each variant uses the same feature embedding and decoder described in Section 3.1.2, swapping only the backbone that processes temporal sequences. GRU achieves the best accuracy across all metrics at the lowest computational cost (154M FLOPs, 565 FPS). LSTM provides the closest accuracy with comparable speed, making it a viable alternative when GRU operators are unavailable (Section 4.2). The Transformer and FastViT [64] lag behind on all accuracy metrics despite higher computational cost.

Table 4: Backbone architecture comparison (all 3-layer).

Backbone	FLOPs	FPS	MPJRE↓	MPJPE↓	MPJVE↓
Transformer	171M	308	4.83	6.94	29.55
FastViT [64]	1027M	39.1	5.49	7.65	29.01
LSTM	190M	542.3	4.43	5.75	22.95
GRU (ours)	154M	565.0	4.19	5.40	21.72

Input robustness. FlowAvatar stays usable under severe tracking loss: with only the head and a single hand available at test time (the model trained with 40% hand-masking augmentation), MPJRE rises modestly from 4.19° to 5.29°.

5.2 Texture Generation

Our texture pipeline adds two design choices to standard differentiable rendering [35]: face alignment and super-resolution upscaling. Because personalized appearance is a supporting component rather than a core contribution, we evaluate these as an internal ablation of our own design, not a benchmark against dedicated avatar methods. We ablate both on the People Snapshot dataset [4] (four sequences, male/female-3/4-casual), holding out one-third of frames for testing with SMPL-X parameters from PyMAF-X rather than the dataset’s,

Table 5: Ablation study on face alignment and upscaling.

Condition	PSNR↑	SSIM↑	LPIPS↓
Baseline	25.289	0.9456	0.0644
+ Face align	25.289	0.9456	0.0640
+ Upscaling	25.291	0.9473	0.0639
+ Both	25.294	0.9478	0.0630

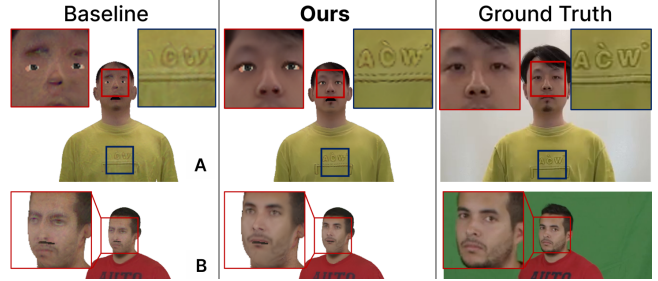


Figure 5: Texture ablation on (A) our capture data and (B) People Snapshot dataset. Face alignment and upscaling improve facial detail.

to simulate real-world conditions. Learned textures are rendered onto the test poses and compared to ground-truth RGB using three standard metrics: PSNR, SSIM, and LPIPS [71].

Table 5 shows both steps help incrementally. The gains are small only because they average over the full frame, where the face is a small share of pixels; on the face itself the improvement is clearly visible (Figure 5): face alignment corrects eye and mouth registration errors from monocular pose, and upscaling recovers detail lost at 512×512 . With the pose model and texture pipeline validated, we next examine how these components shape the user’s experience.

6 USER STUDIES

Having established that FlowAvatar integrates body, hand, face, and eye tracking into a single runtime with competitive, temporally stable pose, we now ask which of these components actually shape the avatar experience, a comparison that prior siloed systems could not support. We investigate two research questions:

- **RQ1:** does tracking completeness or personalized appearance more strongly shape *third-person* perceived motion quality?
- **RQ2:** how do tracking completeness and personalized appearance each shape *first-person* embodiment (ownership, agency, and self-location)?

We address RQ1 with a third-person motion quality study (Section 6.1) and RQ2 with a first-person embodiment study (Section 6.2), each isolating tracking completeness from personalized look-alike texture. Both run in the streaming configuration to log per-frame data under constant conditions.

6.1 Third-Person Motion Quality Study (N=12)

We examined how four avatar representations shape third-person perception of motion quality.

6.1.1 Participants and Procedure

We recruited 12 participants (7M/5F, aged 21–32, $M = 25.17$, $SD = 3.72$) from our university community, all with prior XR experience. To hold motion identical across representations, an experimenter performed each sequence once on a Quest Pro in the streaming configuration. We logged the raw head and hand tracking input together with the model’s predicted full-body pose, then

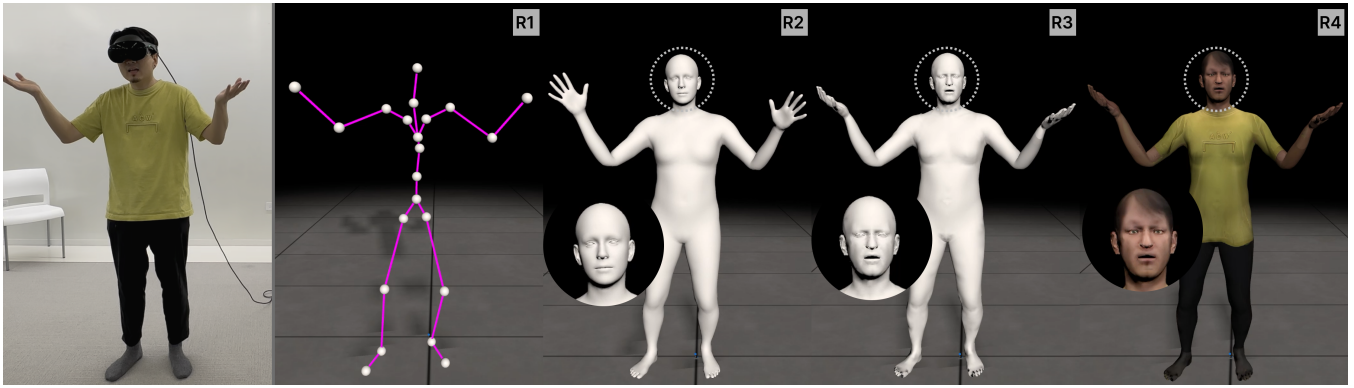


Figure 6: Example sequence (Anger) from the third-person motion quality study: ground truth, then representations (R1) Stick Figure, (R2) Minimal Body (no hands/face), (R3) Base Body (no texture), and (R4) Photo-real (personalized look-alike texture).

Table 6: Motion sequences and evaluation dimensions (the four avatar representations R1–R4 are shown in Figure 6).

Motion Sequences	
S1: Greeting	Waving, smiling, clapping — upper body gestures in social interaction
S2: Anger	Frustrated gestures, angry expressions, assertive locomotion — emotional expressivity
S3: Joy	Jumping, smiling, rhythmic movement — whole-body motion with subtle expressions
S4: Exercise	Squatting, jumping, boxing — dynamic full-body tracking under high intensity
S5: Workplace	Sitting/standing transitions, desktop tasks — fine hand movements and body transitions
Evaluation Dimensions	
Q1	Style & Character: How accurately does this representation capture the style, character, and expression of the original motion?
Q2	Motion Fidelity: How accurately does this representation capture the overall movement qualities?
Q3	Communication: How effectively could you understand what the person was trying to communicate?
Q4	Interaction Readiness: How confidently could you interact with this avatar based on its movements and expressions?
Q5	Upper Body: How effectively does this representation capture upper body movements (head, torso, arms, hands)?
Q6	Lower Body: How effectively does this representation capture lower body movements (hips, legs, feet)?

re-rendered this single recording onto all four representations. Participants viewed the resulting clips (Figure 6) in randomized order and rated each on six 7-point Likert dimensions (−3 to +3, adapted from [61]). Sequences and dimensions are listed in Table 6.

6.1.2 Results

Figure 7 shows mean ratings for the four representations across the six evaluation dimensions. Photo-real (R4) received the highest overall ratings ($M = 1.19$, $SD = 1.09$), followed by Base Body (R3; $M = 0.92$, $SD = 1.11$), while Minimal Body (R2; $M = -0.07$, $SD = 1.27$) and Stick Figure (R1; $M = 0.00$, $SD = 1.60$) scored substantially lower. Friedman tests confirmed significant main effects on all six dimensions (all $p < 0.003$, Kendall’s W 0.42–0.54), despite identical underlying motion. Adding body geometry alone (R1→R2) did not improve ratings: a body mesh without articulated face and hand tracking conveys little beyond a skeleton, so we focus the contrasts on tracking completeness (R2→R3) and texture

Table 7: Motion quality: paired Wilcoxon results for the two key contrasts across Q1–Q6. r is the rank-biserial effect size, CI the 95% bootstrap CI of the paired mean difference. Exact- p markers: $^{\dagger} p < .10$, $^* p < .05$, $^{**} p < .01$, $^{***} p < .001$; bold p_{adj} survives Holm correction.

	+Tracking (R2→R3)		+Texture (R3→R4)	
	p (p_{adj})	r [CI]	p (p_{adj})	r [CI]
Q1	$< .001^{***}$ (.004)	+0.97 [+0.5, +1.1]	.098 [†] (.488)	+0.58 [−0.0, +0.8]
Q2	$< .001^{***}$ (.001)	+1.00 [+0.6, +1.5]	.512 (.948)	+0.25 [−0.2, +0.4]
Q3	$< .001^{***}$ (.001)	+1.00 [+0.8, +1.7]	.316 (.948)	+0.33 [−0.1, +0.5]
Q4	.014 [*] (.014)	+0.85 [+0.3, +1.1]	.004 ^{**} (.023)	+0.96 [+0.3, +1.0]
Q5	.002 ^{**} (.004)	+1.00 [+0.7, +1.8]	.152 (.609)	+0.56 [−0.0, +0.3]
Q6	$< .001^{***}$ (.004)	+1.00 [+0.6, +1.3]	.439 (.948)	+0.29 [−0.2, +0.5]

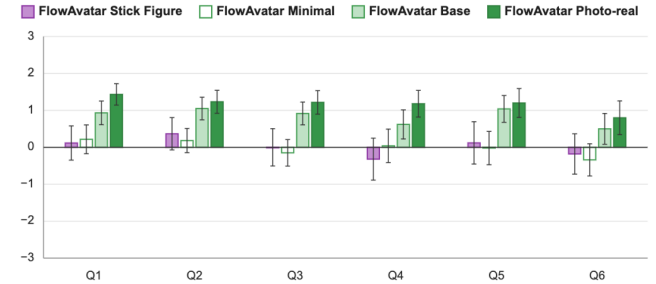


Figure 7: Third-person motion quality study: Mean ratings with standard error for four avatar representations across five motion sequences and six evaluation dimensions.

(R3→R4).

We compared conditions with Wilcoxon signed-rank tests, Holm-corrected across the six dimensions (Table 7). Adding face and hand tracking (R2→R3) improved all six dimensions (all $p_{adj} \leq 0.014$) with large, consistent effect sizes (rank-biserial $r \geq 0.85$, all CIs excluding zero), whereas adding personalized texture (R3→R4) reached significance only for Interaction Readiness (Q4, $p_{adj} = 0.023$).

6.2 First-Person Embodiment Study (N=10)

We assessed FlowAvatar’s effectiveness in creating first-person embodiment, specifically the sense of ownership and agency over one’s virtual avatar.



Figure 8: Embodiment study: (A) participant’s pose; avatar representations: (B) FlowAvatar Base Body; (C) Meta Avatar; (D) FlowAvatar Photo-real (personalized look-alike texture). (E) Five participants’ avatars during study.

6.2.1 Participants and Procedure

We recruited 10 new participants (6M/4F, mean age 30.1, $SD = 9.46$), all with prior XR experience. Sessions averaged 53 minutes. Each first created a personalized avatar with our web tool and a cartoon avatar in Meta Horizon [44]. All conditions ran in the same Unity scene, a virtual mirror (Figure 8) in which participants performed the same free-form movements (walking, jumping, gestures, and facial expressions), so only the avatar differed: two were driven by FlowAvatar, Base Body (full tracking, no texture) and Photo-real (full system with personalized texture), while the third loaded the participant’s Meta Horizon avatar (upper body only) after signing in to Meta. Conditions were shown in varying order across participants, with participants blind to the driving system. The Photo-real vs. Base Body comparison isolates texture with tracking held constant, whereas Photo-real vs. Meta Avatar additionally varies tracking and visual style, serving as an external consumer baseline rather than a controlled contrast. Participants rated each condition on the CoolMoves questionnaire [2, 50]: ownership (± 15), agency (± 12), location (± 9), appearance (± 6), response (± 6), and total embodiment (± 3). After each session, a ~ 10 -minute semi-structured interview was coded by two researchers into recurring themes (self-recognition, motion coherence, and tracking predictability), illustrated with representative quotes below.

6.2.2 Results

Figure 9 shows CoolMoves [2, 50] subscale scores across all three conditions. Photo-real scored numerically highest on all six dimensions and was the only condition achieving positive total embodiment. A Friedman test showed a significant main effect for Appearance ($\chi^2(2) = 6.74$, $p = 0.034$, $W = 0.337$) and a marginal trend for Agency ($\chi^2(2) = 5.03$, $p = 0.081$).

Ownership. Photo-real achieved the highest ownership scores ($M = +3.60$, $SD = 7.04$), with participants noting personal recognition: “Since FlowAvatar has the same outfit as mine, I think FlowAvatar is my avatar” [P-004]. Meta Avatar showed moderate ownership ($M = +2.40$, $SD = 6.72$), though participants found it unconvincing: “I don’t think the cartoon looks like me” [P-002]. Base Body scored lowest ($M = -1.50$, $SD = 6.92$), lacking personalization despite identical motion to Photo-real. The Photo-real vs. Base Body difference showed a marginal trend ($p = 0.062$).

Table 8: Embodiment ($N=10$, exploratory): paired Wilcoxon results. Photo vs. Base isolates texture (tracking held constant). Photo vs. Meta additionally varies tracking and style. r is rank-biserial, CI the 95% bootstrap CI of the paired mean difference. Exact- p markers: $^\dagger p < .10$, $^* p < .05$; no comparison survives Holm correction (p_{adj}).

Subscale	Photo vs. Base		Photo vs. Meta	
	$p(p_{adj})$	r [CI]	$p(p_{adj})$	r [CI]
Ownership	.062 [†] (.250)	+0.67 [+0.7, +9.2]	.668 (1.000)	+0.18 [−4.6, +6.6]
Agency	.125 (.250)	+0.64 [+0.0, +3.1]	.043 [*] (.258)	+0.73 [+0.8, +7.3]
Location	.102 (.250)	+0.67 [+0.1, +3.9]	.109 (.547)	+0.64 [+0.2, +6.4]
Appearance	.031 [*] (.164)	+1.00 [+0.9, +5.0]	1.000 (1.000)	−0.04 [−1.3, +2.4]
Response	.078 [†] (.250)	+0.82 [+0.3, +3.0]	.500 (1.000)	+0.67 [−0.2, +2.8]
Total	.027 [*] (.164)	+0.78 [+0.3, +1.2]	.375 (1.000)	+0.35 [−0.0, +1.5]

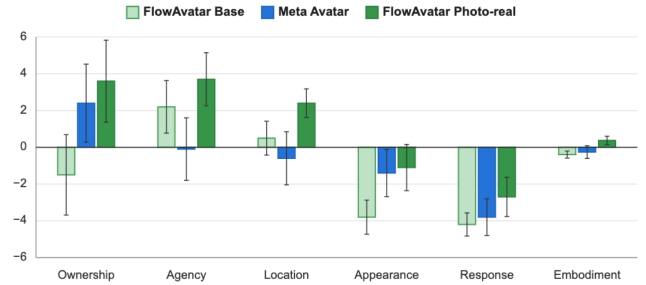


Figure 9: First-person embodiment study: CoolMoves [2, 50] subscale scores (centered at 0) for three avatar conditions. Photo-real was the only condition with positive Total Embodiment.

Agency. Photo-real ($M = +3.70$, $SD = 4.57$) substantially exceeded Meta Avatar ($M = -0.10$, $SD = 5.38$), reaching significance ($p = 0.043$). Participants specifically mentioned motion capabilities: “What surprised me is that the FlowAvatar’s jump and squat are nice” [P-006]. A key differentiator was tracking behavior: “When I raise my hands up...Meta ones will fall down, which makes me feel scary because I did not do that” [P-007], highlighting how FlowAvatar’s approach to handling tracking loss created a more predictable experience. Base Body scored moderately ($M = +2.20$, $SD = 4.52$). Meta Avatar’s lack of lower-body animation was a significant limitation for agency.

Location. Photo-real scored highest for body location ($M = +2.40$, $SD = 2.46$), with integrated tracking creating spatial presence: “All my tracking, hands, face are pretty nice in FlowAvatar, with these all together, I can see in mirror it was me” [P-008]. Base Body received moderate positive scores ($M = +0.50$, $SD = 2.92$), as one participant noted “I know the white body is not me, it only has my motion data” [P-010], while Meta Avatar received lower location scores ($M = -0.60$, $SD = 4.58$), with participants reporting disconnection between their physical movements and the avatar’s representation.

Appearance and Response. Photo-real ($M = -1.10$, $SD = 3.98$) scored higher than Base Body ($M = -3.80$, $SD = 2.94$) on Appearance (uncorrected $p = 0.031$, Holm-corrected $p_{adj} = 0.164$), the only dimension with a significant Friedman main effect. Meta Avatar scored between the two ($M = -1.40$, $SD = 4.06$). All three conditions received negative appearance scores. For Response, Photo-real scored highest ($M = -2.70$, $SD = 3.37$) followed by Meta Avatar ($M = -3.80$, $SD = 3.16$) and Base Body ($M = -4.20$, $SD = 1.99$), with a marginal Photo-real vs. Base Body trend ($p = 0.078$). Participants generally did not feel their real bodies were affected by virtual avatar interactions.

Total Embodiment. Photo-real ($M = +0.37$, $SD = 0.73$) was the only condition achieving positive total embodiment, scoring higher than Base Body ($M = -0.40$, $SD = 0.61$, uncorrected $p = 0.027$, $p_{adj} = 0.164$). Meta Avatar ($M = -0.26$, $SD = 1.10$) also scored below neutral.

We also compared conditions pairwise with Wilcoxon signed-rank tests, Holm-corrected across the six subscales and treated as exploratory (Table 8). The Photo-real vs. Base Body comparison (isolating texture) reached significance on Appearance ($p = .031$) and Total Embodiment ($p = .027$), with marginal trends on Ownership ($p = .062$) and Response ($p = .078$), whereas the Photo-real vs. Meta Avatar comparison reached significance only on Agency ($p = .043$). These effects carried large effect sizes (rank-biserial $r \geq 0.64$), but none survived Holm correction at $N = 10$.

7 DISCUSSION

Returning to our two research questions, both studies point to a consistent division of labor between tracking and appearance. In the motion quality study (RQ1), adding face and hand tracking improved all six evaluation dimensions (Holm-corrected, all $p_{adj} \leq 0.014$), whereas adding texture reached significance on only one, so tracking completeness dominates third-person motion quality. The effect was strongest for emotional content: in the Anger sequence, face/hand tracking raised Style and Character scores from $M = 0.00$ – 0.08 to $M = 1.38$ – 1.58 , indicating that articulated hands and faces carry emotional intent that body pose alone conveys poorly. In the embodiment study (RQ2), FlowAvatar’s tracking was associated with greater agency than Meta Avatar ($p = .043$, uncorrected) and texture with ownership and self-recognition. We read the agency effect as suggestive, given the confounded Meta comparison. Participants consistently valued motion coherence, as “For the avatar movement, it has more natural movement” [P-002], and predictable tracking behavior, with P-007 noting that Meta Avatar’s hands would unexpectedly drop when raised overhead.

This user-perceived quality follows from the model’s design. The comprehensive temporal losses (Section 3.1.3) are the single most impactful loss category in our ablation, reducing jitter by 37% (Table 3), and this temporal consistency plausibly underlies the reliable agency participants reported: predictable full-body behavior mattered more for felt agency than visual fidelity. More broadly, for sparse egocentric tracking, pairing comprehensive spatial and temporal losses with a lightweight recurrent architecture yields the best temporal consistency in our benchmark (Table 4).

Personalized texture serves a complementary but distinct role. It scored higher on Appearance ($p = .031$) and Total Embodiment ($p = .027$) than Base Body, though neither effect survives Holm correction at this small sample. Participants valued self-recognition: “I prefer the photorealistic one, compared to the cartoon one, it looks more like me” [P-009]. However, uncanny valley effects emerged: “It is like someone with your clothes but looks a bit different than you, but still, it looks pretty close to me already” [P-003]. Several participants preferred cartoon avatars in social contexts: “I still think virtual world should be something virtual, you can’t bring the ‘real’ me there” [P-010]. We also observed that current limitations in hair modeling and facial geometry may affect avatar quality unevenly across users, as “it can’t have better solution to show my hair, which makes me ugly” [P-004]. These reactions suggest that the value of personalized look-alike texture depends on application context and avatar quality, and that equitable representation remains an open challenge. Practically, when compute-constrained, these preliminary findings suggest prioritizing tracking over rendering fidelity: in our studies, complete tracking of body, face, and hands had the broadest impact on motion quality and agency, while texture added value mainly through self-recognition and ownership. Notably, these interactions between tracking and appearance only become visible through integrated evaluation, reinforcing the value of holistic sys-

tem assessment over optimizing avatar subsystems independently.

8 LIMITATIONS AND FUTURE WORK

Our pose estimation generalizes well across diverse movements, though it occasionally faces challenges with poses not well-represented in training data. The current implementation treats environmental interactions as soft constraints through loss functions rather than hard physical constraints, achieving plausible results at runtime. Future work could extend beyond flat ground planes to include full-body environment contact [66], object interactions [9], and additional sensor fusion (e.g., IMUs) for improved lower-body estimation.

The SMPL-X mesh constrains visual fidelity: Unity’s implementation reduces the blendshape space from 100 to 10 expressions, and the model cannot represent hair, clothing dynamics, or changing lighting, with no separation between skin and fabric. External-sensor capture would reach higher photorealism, a fidelity we deliberately trade for a smartphone-only, consumer-accessible enrollment pipeline. Hybrid approaches with codec avatars [41] or dynamic rendering [52], and emerging Gaussian Splatting [26], could address these. Recent mobile-ready Gaussian avatar distillation [27] suggests such upgrades are becoming feasible on consumer hardware.

Our on-device configuration was not evaluated perceptually. Future work should assess whether its quality difference from the streaming configuration, and the tracker-bound 97 ms motion delay (Section 4.2), are noticeable in practice.

The user studies are preliminary. With modest samples ($N=12$, $N=10$) and no a-priori power analysis, we report effect sizes and confidence intervals throughout, and several embodiment effects (such as Photo-real over Base Body on Appearance and Total Embodiment) hold only before multiple-comparison correction. The Meta Avatar condition is an external baseline, not a controlled contrast: its closed interface varies tracking, rendering, and lower-body animation together. In the motion quality study, the Q1 item bundled style, character, and expression into one rating, limiting per-component attribution, though the tracking effect held across all six items. Larger, appropriately powered studies are needed to confirm these effects.

Finally, FlowAvatar integrates multimodal motion signals (head, hand, face, gaze, full-body pose) that can reveal sensitive personal attributes, with risks compounding across modalities [47]. Our on-device runtime mitigates this by keeping raw motion on the headset rather than streaming it; privacy-preserving representation and storage remain important future work.

9 CONCLUSION

FlowAvatar demonstrates that a unified avatar system connecting personalized look-alike appearance generation to real-time full-body pose, hand articulation, facial expression, and eye gaze tracking is achievable on a standalone consumer headset without additional instrumentation. A pose estimation framework combining body-part encoding, unified kinematic decoding, and comprehensive spatial and temporal losses achieves competitive spatial accuracy with state-of-the-art temporal consistency on the AMASS benchmark for this sparse input regime. Two user studies provide converging evidence that tracking completeness and personalized appearance serve complementary roles: tracking completeness drives perceived motion quality and is associated with agency, personalized texture with ownership and self-recognition, and only their combination achieves positive total embodiment. We release the complete pipeline as open source at <https://github.com/SPICeXLAB/FlowAvatar>, lowering the barrier to embodied avatars on consumer hardware and inviting the community to build on these findings.

REFERENCES

- [1] K. Ahuja, C. Harrison, M. Goel, and R. Xiao. Mecap: Whole-body digitization for low-cost VR/AR headsets. In *Proceedings of the 32nd*

- Annual ACM Symposium on User Interface Software and Technology*, pp. 453–462, 2019. 2
- [2] K. Ahuja, E. Ofek, M. Gonzalez-Franco, C. Holz, and A. D. Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021. 1, 8
 - [3] S. Aliakbarian, F. Saleh, D. Collier, P. Cameron, and D. Cosker. Hmd-nemo: Online 3D avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9622–9631, 2023. 2
 - [4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8387–8397, 2018. doi: 10.1109/CVPR.2018.00875 6
 - [5] Apple Inc. ARKit: Tracking and visualizing faces (ARFaceAnchor blend shapes). <https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation>, 2024. 4
 - [6] Apple Inc. Set up your Persona (beta) on Apple Vision Pro, 2024. 1, 2
 - [7] T. Baltrušaitis, P. Robinson, and L.-P. Morency. OpenFace: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–10. IEEE, 2016. 2
 - [8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pp. 561–578. Springer, 2016. 2
 - [9] S. Brahmabhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 361–378. Springer, 2020. 9
 - [10] J. Chen, J. Wang, Y. Zhang, R. Pandey, T. Beeler, M. Habermann et al. Egoavatar: Egocentric view-driven and photorealistic full-body avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024. 2
 - [11] X. Chen, Y. Zhu, Y. Li, B. Fu, L. Sun, Y. Shan et al. Robust human matting via semantic guidance. In *Proceedings of the Asian Conference on Computer Vision*, pp. 2984–2999, 2022. 4
 - [12] B. A. Cohn, A. Maselli, E. Ofek, and M. Gonzalez-Franco. Snapmove: Movement projection mapping in virtual reality. In *2020 IEEE international conference on artificial intelligence and virtual reality (AIVR)*, pp. 74–81. IEEE, 2020. 1, 6
 - [13] M. Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020. 4
 - [14] P. Dai, Y. Zhang, T. Liu, Z. Fan, T. Du, Z. Su et al. Hmd-pose: On-device real-time human motion tracking from scalable sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 874–884, 2024. 2, 3, 4, 5
 - [15] H. G. Debarba, S. Chague, and C. Charbonnier. On the plausibility of virtual body animation features in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi: 10.1109/TVCG.2020.3025175 3
 - [16] N. Devrio and C. Harrison. Discoband: Multiview depth-sensing smart-watch strap for hand, body and environment tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–13, 2022. 2
 - [17] Y. Du, R. Kips, A. Pumarola, S. Starke, A. Thabet, and A. Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 481–490, 2023. 2, 3, 5, 6
 - [18] J. C. Eubanks, A. G. Moore, P. A. Fishwick, and R. P. McMahan. The effects of body tracking fidelity on embodiment of an inverse-kinematic avatar for male participants. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2020. doi: 10.1109/ISMAR50242.2020.00025 1, 3
 - [19] M. L. Fiedler, C. Merz, S. Schach, J. Tschanter, and M. E. Latoschik. Am I still me? visual congruence across reality-virtuality and avatar appearance in shaping self-perception and behavior. *IEEE Transactions on Visualization and Computer Graphics*, 2026. doi: 10.1109/TVCG.2026.3679890 3
 - [20] R. Fribourg, F. Argelaguet, A. Lécuyer, and L. Hoyet. Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view. *IEEE transactions on visualization and computer graphics*, 26(5):2062–2072, 2020. 1, 3
 - [21] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14783–14794, 2023. 2
 - [22] M. Gonzalez-Franco, A. I. Bellido, K. J. Blom, M. Slater, and A. Rodriguez-Fornells. The neurological traces of look-alike avatars. *Frontiers in human neuroscience*, 10:392, 2016. 1, 3
 - [23] M. Gonzalez-Franco, B. Cohn, E. Ofek, D. Burin, and A. Maselli. The self-avatar follower effect in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 18–25. IEEE, 2020. 1
 - [24] M. Gonzalez-Franco, A. Steed, S. Hoogendyk, and E. Ofek. Using facial animation to increase the enfacement illusion and avatar self-identification. *IEEE transactions on visualization and computer graphics*, 26(5):2023–2029, 2020. 1, 3
 - [25] D. Halbhüser, M. Kocur, A. Kalus, K. Angermeyer, V. Schwind, and N. Henze. Understanding the effects of perceived avatar appearance on latency sensitivity in full-body motion-tracked virtual reality. In *Proceedings of Mensch und Computer 2023*. ACM, 2023. doi: 10.1145/3603555.3603580 5
 - [26] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang et al. Gaussianavator: Towards realistic human avatar modeling from a single video via animatable 3D gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 634–644, 2024. 2, 9
 - [27] F. Iandola, S. Pidhorskyi, I. Santesteban, D. Gupta, A. Pahuja, N. Bartolovic et al. Squeezeme: Mobile-ready distillation of gaussian full-body avatars. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025. 2, 9
 - [28] L. Izzouzi and A. Steed. The influence of avatar realism and similarity in shaping group dynamics in collaborative virtual environments. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 604–607. IEEE, 2024. 3
 - [29] J. Jiang, P. Strel, M. Meier, and C. Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pp. 277–294. Springer, 2024. 2, 3, 4, 5, 6
 - [30] J. Jiang, P. Strel, H. Qiu, A. Fender, L. Laich, P. Snape et al. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pp. 443–460. Springer, 2022. 2, 3, 4, 5
 - [31] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2
 - [32] K. Kiltner, R. Groten, and M. Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012. doi: 10.1162/PRES_a_00124 1
 - [33] D. Kim and C. Harrison. Etherpose: Continuous hand pose tracking with wrist-worn antenna impedance characteristic sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–12, 2022. 2
 - [34] A. Kyu, H. Mao, J. Zhu, M. Goel, and K. Ahuja. Eitpose: Wearable and practical electrical impedance tomography for continuous hand pose estimation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2024. 2
 - [35] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 5, 6
 - [36] J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes et al. Project Starline: A high-fidelity telepresence system. *ACM Transactions on Graphics*, 40(6):1–16, 2021. doi: 10.1145/3478513.3480490 2
 - [37] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou et al. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1144–

- [38] Y. Liu, S. Zhang, and M. Gowda. Neuropose: 3D hand pose tracking using EMG wearables. In *Proceedings of the Web Conference 2021*, pp. 1471–1482, 2021. 2
- [39] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. De La Torre et al. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 64–73, 2021. 2
- [40] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019. 5
- [41] J. Martinez, E. Kim, J. Romero, T. Bagautdinov, S. Saito, S.-I. Yu et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 9
- [42] T. Menzel, E. Wolf, S. Wenninger, N. Spinczyk, L. Holderrieth, C. Wienrich et al. Avatars for the masses: Smartphone-based reconstruction of humans for virtual reality. *Frontiers in Virtual Reality*, 6:1583474, 2025. doi: 10.3389/frvir.2025.1583474 3
- [43] Meta Platforms, Inc. Inside-out body tracking (IOBT) for quest 3. Movement SDK for Unity and OpenXR, 2023. 1, 2
- [44] Meta Platforms, Inc. Meta Horizon avatars. <https://www.meta.com/avatars/>, 2024. 4, 8
- [45] V. Molyn, R. Arakawa, M. Goel, C. Harrison, and K. Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2023. 2
- [46] A. Mottelson, A. Muresan, K. Hornbæk, and G. Makransky. A systematic review and meta-analysis of the effectiveness of body ownership illusions in virtual reality. *ACM Transactions on Computer-Human Interaction*, 30(5):1–42, 2023. doi: 10.1145/3590767 3
- [47] V. Nair, C. Rack, W. Guo, R. Wang, S. Li, B. Huang et al. Inferring private personal attributes of virtual reality users from ecologically valid head and hand motion data. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2024. doi: 10.1109/NRW62533.2024.00094 9
- [48] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev et al. Holoportation: Virtual 3D teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pp. 741–754, 2016. 2
- [49] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas et al. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019. 2, 3
- [50] T. C. Peck and M. Gonzalez-Franco. Avatar embodiment: a standardized questionnaire. *Frontiers in Virtual Reality*, 1:575943, 2021. 3, 8
- [51] T. C. Peck, S. Seinfeld, S. M. Aglioti, and M. Slater. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and cognition*, 22(3):779–787, 2013. 3
- [52] S. Saito, G. Schwartz, T. Simon, J. Li, and G. Nam. Relightable Gaussian codec avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 130–141, 2024. 2, 9
- [53] A. Salagean, E. Crellin, M. Parsons, D. Cosker, and D. Stanton Fraser. Meeting your virtual twin: Effects of photorealism and personalization on embodiment, self-identification and perception of self-avatars in virtual reality. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–16, 2023. 3
- [54] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 5
- [55] M. Slater and M. V. Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:74, 2016. 1
- [56] M. Slater, B. Spanlang, and D. Corominas. Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM transactions on graphics (TOG)*, 29(4):1–9, 2010. 3
- [57] M. Slater and S. Wilbur. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, 1997. doi: 10.1162/pres.1997.6.6.603 1
- [58] B. Spanlang, J.-M. Normand, D. Borland, K. Kilteni, E. Giannopoulos, A. Pomés et al. How to build an embodiment lab: achieving body representation illusions in virtual reality. *Frontiers in Robotics and AI*, 1:9, 2014. 1
- [59] A. Steed, Y. Pan, F. Zisch, and W. Steptoe. The impact of a self-avatar on cognitive load in immersive virtual reality. In *2016 IEEE virtual reality (VR)*, pp. 67–76. IEEE, 2016. 3
- [60] D. Svitov, P. Morerio, L. Agapito, and A. Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 4051–4068, 2024. 4, 5
- [61] Y. Tao, C. Y. Wang, A. D. Wilson, E. Ofek, and M. Gonzalez-Franco. Embodying physics-aware avatars in virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2023. 3, 7
- [62] J. Tschanter, C. Merz, M. L. Fiedler, C. Wienrich, and M. E. Latoschik. Use case matters: Comparing the user experience and task performance across tasks for embodied interaction in VR. *IEEE Transactions on Visualization and Computer Graphics*, 2026. doi: 10.1109/TVCG.2026.3680577 3
- [63] T. Van Wouwe, S. Lee, A. Falisse, S. Delp, and C. K. Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2513–2523, 2024. 2
- [64] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5785–5795, 2023. 6
- [65] X. Wang, L. Xie, C. Dong, and Y. Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021. 5
- [66] A. Winkler, J. Won, and Y. Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–8, 2022. 9
- [67] E. Wu, Y. Yuan, H.-S. Yeo, A. Quigley, H. Koike, and K. M. Kitani. Back-hand-pose: 3D hand pose estimation for a wrist-worn camera via dorsum deformation network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 1147–1160, 2020. 2
- [68] V. Xu, C. Gao, H. Hoffmann, and K. Ahuja. Mobileposer: Real-time full-body pose estimation and 3D human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–11, 2024. 2
- [69] X. Yi, Y. Zhou, and F. Xu. Transpose: Real-time 3D human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)*, 40(4):1–13, 2021. 2
- [70] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun et al. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023. 4
- [71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6
- [72] X. Zheng, Z. Su, C. Wen, Z. Xue, and X. Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14678–14688, 2023. 2, 3, 5
- [73] T. Zhi, C. Lassner, T. Tung, C. Stoll, S. G. Narasimhan, and M. Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 492–509. Springer, 2020. 2
- [74] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15475–15484, 2023. 5