








VRGaussianAvatar: Integrating 3D Gaussian Avatars into VR

Hail Song , Boram Yoon , Seokhwan Yang , Seoyoung Kang ,
Hyunjeong Kim , Henning Metzmacher , and Woontack Woo 

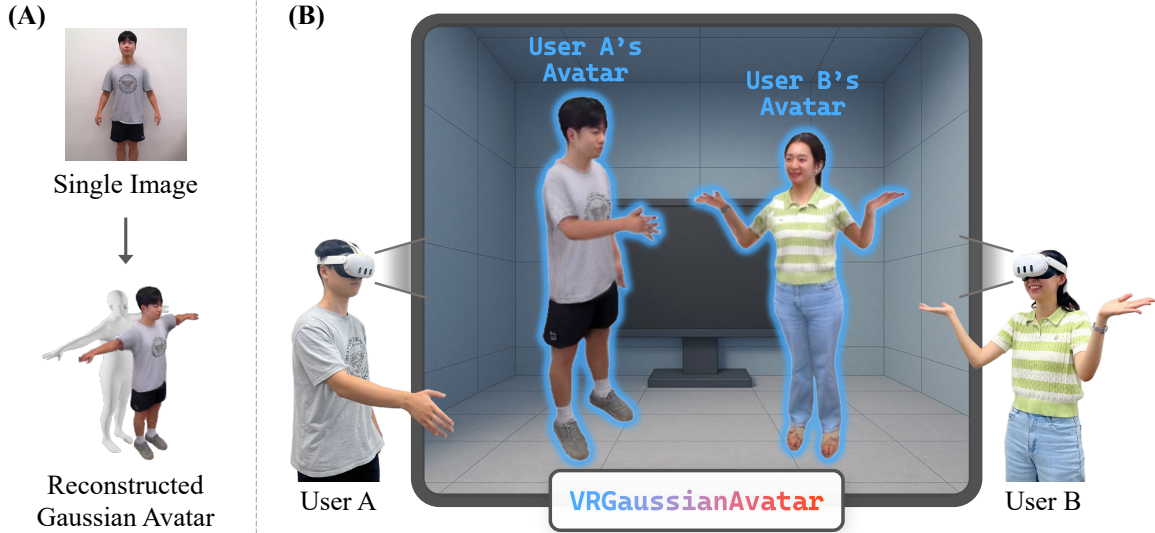


Fig. 1: We introduce **VRGaussianAvatar**, an integrated VR system for controllable full-body 3DGS avatar representation. (A) A single input image is used to reconstruct a full-body 3D Gaussian Splatting avatar. (B) The reconstructed avatar supports real-time rendering and full controllability in VR using only the internal sensors of commercial head-mounted displays (HMDs).

Abstract—We present VRGaussianAvatar, an integrated system that enables real-time full-body 3D Gaussian Splatting (3DGS) avatars in virtual reality using only head-mounted display (HMD) tracking signals. The system adopts a parallel pipeline with a VR Frontend and a GA Backend. The VR Frontend uses inverse kinematics to estimate full-body pose and streams the resulting pose along with stereo camera parameters to the backend. The GA Backend stereoscopically renders a 3DGS avatar reconstructed from a single image. To improve stereo rendering efficiency, we introduce Binocular Batching, which jointly processes left and right eye views in a single batched pass to reduce redundant computation and support high-resolution VR displays. We evaluate VRGaussianAvatar with quantitative performance tests and a within-subject user study against image- and video-based mesh avatar baselines. Results show that VRGaussianAvatar sustains interactive VR performance and yields higher perceived appearance similarity, embodiment, and plausibility. Source code is available at <https://github.com/hailsong/VRGaussianAvatar>.

Index Terms—Virtual reality, Augmented reality, 3D Gaussian Splatting, Avatar reconstruction, Stereoscopic rendering

1 INTRODUCTION

The representation of photorealistic full-body avatars in Virtual Reality (VR) plays a crucial role in enhancing user immersion and effective communication. Applications such as telepresence and gaming benefit significantly from avatars that faithfully replicate the user’s appearance.

To make avatar creation practical and accessible, existing methods

allow for the creation of mesh-based 3D avatars. These methods reconstruct drivable human models from a single image [3, 10, 20, 38, 59] or from video sequences [1, 23, 48]. While these approaches are highly compatible with existing graphics pipelines and offer strong real-time performance, they often struggle to deliver photorealistic avatar representations for subjects that significantly deviate from template meshes such as SMPL-X [38]. Recently, neural rendering techniques, including Neural Radiance Fields (NeRF) [34] and 3D Gaussian Splatting (3DGS) [21], have achieved remarkable success across diverse 3D visualization tasks. These methods have been applied not only to novel view synthesis but also to object reconstruction [24, 49, 50] and avatar creation [4, 12, 17]. In particular, 3DGS represents a 3D scene as a collection of colored Gaussian splats, enabling real-time photorealistic rendering and demonstrating strong performance in both face [13, 52, 65] and body avatar generation [35, 42, 45].

In VR, several studies [18, 53] have proposed techniques that leverage 3DGS for immersive scene representation. Kim et al. [22] conducted quantitative evaluations and user studies to examine the effectiveness of 3DGS-based object representation in VR, confirming that 3DGS is effective for realistic object rendering in virtual environments. VOODOO XP [51] was the first to propose a VR telepresence application using a 3DGS-based head avatar animated using the internal sensors of head-mounted displays (HMDs).

- Hail Song is with KAIST UVR Lab. E-mail: hail96@kaist.ac.kr
- Boram Yoon is with KAIST KI-ITC ARRC. E-mail: boram.yoon1206@kaist.ac.kr
- Seokhwan Yang is with KAIST UVR Lab. E-mail: ysshwan147@kaist.ac.kr
- Seoyoung Kang is with KAIST UVR Lab. E-mail: sy1009kang@kaist.ac.kr
- Hyunjeong Kim is with Hansung University and ETH Zurich, Game Technology Center. E-mail: hcjhkim@gmail.com
- Henning Metzmacher is with ETH Zurich, Game Technology Center. E-mail: henning.metzmacher@inf.ethz.ch
- Woontack Woo is with KAIST UVR Lab and KAIST KI-ITC ARRC. E-mail: wwoo@kaist.ac.kr

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxxx

However, several challenges have prevented 3DGS-based full-body avatars from being deployed in VR communication systems. In addition to the significant hardware demands of 3DGS visualization, VR requires stereoscopic rendering which imposes even higher computational demands. Moreover, 3DGS pipelines have limited compatibility with conventional VR rendering pipelines. Real-time animation of 3DGS avatars remains challenging, and existing systems do not directly utilize user motion captured from the internal sensors of HMDs. These limitations have hindered the integration of 3DGS avatars into VR.

To overcome these limitations, we present **VRGaussianAvatar**, an integrated system that enables real-time user representation in VR through a 3DGS avatar reconstructed from a single image and controllable using only the internal sensors of a commercial HMD. Our system consists of two subsystems: a VR Frontend, which processes sensor data from a commercial HMD and applies inverse kinematics for avatar animation, and a GA Backend (Gaussian Avatar Backend), which performs real-time photorealistic rendering with 3DGS. To support stereoscopic VR rendering, we introduce **Binocular Batching**, which achieves efficient and consistent rendering for both eyes. The system builds on standard components (e.g., off-the-shelf inverse kinematics and recent single-image 3DGS avatar reconstruction models), while our main contributions are the VR streaming integration, Binocular Batching, and a perceptual user study of 3DGS avatars in VR.

We validate our system through user studies and quantitative evaluations comparing VRGaussianAvatar with mesh-based avatars reconstructed from images and videos. These mesh avatars were built upon the parametric whole-body model SMPL-X [38] and were designed to be controlled solely through the internal sensors of the HMD, similar to ours. For the user study, subjects followed instructions to control both baseline and our method’s avatars, modeled after themselves and others. They then completed questionnaires assessing virtual embodiment and plausibility. The results show that our method provides stronger self-embodiment than mesh-based baselines. Quantitative evaluations further confirm that avatars produced by our approach resemble the user’s real appearance more closely.

In summary, our contributions are as follows:

- **VRGaussianAvatar**: an integrated VR streaming system for stereoscopic rendering of a 3DGS full-body avatar driven by HMD-only tracking.
- **Binocular Batching**: an efficient stereoscopic rendering strategy that jointly renders the left and right eye views in a single batched pass for 3DGS avatars in VR.
- A quantitative evaluation and perceptual user study showing improved perceived appearance similarity, embodiment, and plausibility over mesh-based avatar baselines under the same HMD-only control setting.

2 RELATED WORK

2.1 Avatar Mediated Communication

In VR environments, virtual embodiment and avatar plausibility during communication are both crucial. For effective VR communication, it is important not only that individuals identify with their own avatars through virtual embodiment, but also that avatars are perceived as plausible in appearance and behavior. Previous studies have shown that visual appearance plays a key role in virtual embodiment and more realistic appearances resembling actual humans significantly enhance embodiment [56]. To achieve high levels of plausibility in VR communication, avatar realism is crucial. Studies have concluded that high-quality avatars enhance plausibility-related judgments and user comfort. For example, Yoon et al. [60] demonstrated that realistic full-body avatars are advantageous regarding virtual embodiment and plausibility. Similarly, Cho et al. [6] showed that realistic volumetric avatars yield higher perceived plausibility than pre-scanned avatars when performing both dynamic and static tasks. Moreover, Kaiser et al. [19] reported through quantitative and qualitative experiments that higher avatar realism generally improves plausibility and comfort.

In this work, we present a VR avatar representation system that leverages a full-body avatar based on 3D Gaussian Splatting (3DGS), which

has recently shown strong performance in realistic 3D human body model representation. Furthermore, through user studies and statistical analysis, we confirm that the proposed Gaussian avatar representation system contributes to increasing users’ virtual embodiment and avatar plausibility.

2.2 3D Avatar Reconstruction

The generation of 3D avatars resembling users has been a critical topic in the areas of computer vision, computer graphics, and virtual reality. For full-body avatar reconstruction, the parametric body model SMPL-X [38] has been widely adopted, as it encodes body shape and facial expressions into latent features and enables pose-dependent 3D representations and animations driven by body, hand, and facial configurations. Using SMPL-X as a template mesh, several studies have explored approaches to reconstruct mesh and texture maps from videos [1, 48, 67] or images [28, 38].

The advancement of neural rendering techniques, represented by neural radiance fields (NeRF) [34], has demonstrated strong performance in generating photorealistic full-body avatars from videos [4, 17, 61, 66] or images [14]. However, the high computational cost and long processing times of these methods limit their applicability to real-time avatar representation in VR environments. To achieve realistic 3D representations with efficient rendering, 3DGS was introduced as a neural rendering technique designed for both fidelity and speed. 3DGS enables fast and photorealistic avatar representation from videos [27, 35, 36, 39, 41] or image [32, 42, 43, 45] and has also shown superior performance in head model reconstruction [5, 40, 46, 51, 52, 64, 65] compared to NeRF-based approaches.

In particular, the Large Human Model (LHM) [42] achieves one-shot avatar reconstruction in just a few seconds while supporting real-time rendering on consumer-grade GPUs. We adopt LHM [42] as an off-the-shelf backbone to reconstruct 3D Gaussian avatars, and focus on integrating it into a practical VR streaming and stereoscopic rendering system.

2.3 VR Avatar Representation Using Neural Rendering

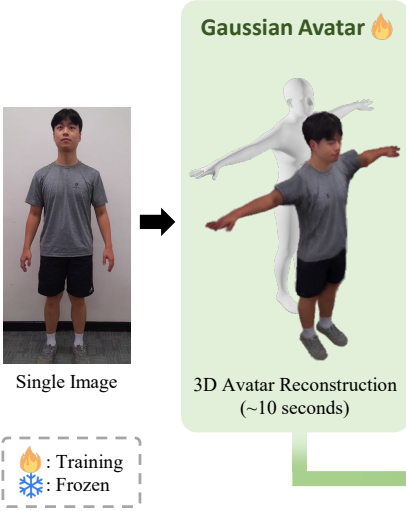
There have been several studies that explored the use of neural rendering for photorealistic scene and object representation in VR. Immersive-NGP [25], which applied the hash-table-based instant-NGP [37] to accelerate NeRF rendering to near real-time, was the first work to demonstrate an application of neural rendering in VR. VR-GS [18] further leveraged 3DGS, which enables both faster rendering and more realistic 3D representations, to achieve real-time VR rendering and introduced physics-based interactions. However, these works have been limited in that they cannot be directly applied to neural rendering-based VR avatar representation, which requires real-time rendering of dynamically moving neural primitives.

To address this, SqueezeMe [16] reduced the computational overhead of animation and 3DGS rendering, making real-time full-body avatar representation in VR. For head avatars, VOODOO XP [51] demonstrated the animation and rendering of a 3DGS head avatar driven by the internal sensors of HMD. Nevertheless, no prior work has realized a full-body avatar representation using only the limited information from HMD-based body tracking and head tracking. Motivated by these gaps, we integrate a full-body 3DGS avatar into a VR streaming pipeline and evaluate its perceptual impact in VR.

3 METHOD

VRGaussianAvatar is structured as a parallel framework with a **VR Frontend** and a **GA Backend**. The VR Frontend estimates full-body pose from HMD sensors and provides per-frame camera parameters. The GA Backend animates and renders the pre-reconstructed 3DGS avatar, incorporating a **Binocular Batching** strategy for efficient stereoscopic rendering. In our reference setup, the VR Frontend and GA Backend communicate over localhost to minimize network overhead. Figure 2 presents the overall system architecture.

Avatar Reconstruction (Offline)



Runtime Process (Online)

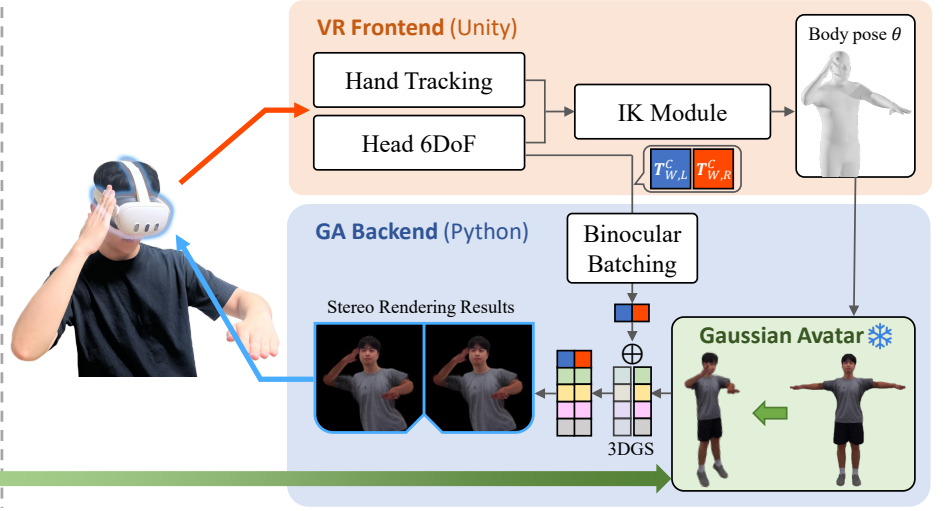


Fig. 2: System diagram of the proposed method, **VRGaussianAvatar**. *Avatar reconstruction phase (left)*: Given a single input image, the Gaussian Avatar Module reconstructs a 3D full-body avatar. *Runtime process (right)*: The **VR Frontend** and **GA Backend** operate in parallel to animate and render the avatar in real time. The VR Frontend estimates a SMPL-X-compatible full-body pose from head 6DoF and hand-tracking signals using the IK Module. The GA Backend animates and renders the reconstructed Gaussian avatar, and applies the proposed **Binocular Batching** with both world-to-camera matrices ($\mathbf{T}_{W,L}^C, \mathbf{T}_{W,R}^C$) to enable efficient real-time stereoscopic rendering.

3.1 VR Frontend

Inputs and Objective. The VR Frontend uses only device-native tracking signals from a commercial HMD: head and hand poses (position and orientation). Its goal is to estimate a full-body pose at interactive rates from these sparse cues and to stream per-frame stereo camera parameters to the GA Backend.

Inverse Kinematics Module (IK Module). The VR Frontend reconstructs a full-body pose using an off-the-shelf IK solver. The head pose provides the root constraint via a calibrated head-to-root offset. When available, left/right hand tracking is used as end-effector constraints to stabilize upper-limb articulation; otherwise, the solver falls back to a head-only configuration with posture priors on the torso and shoulders. The solver enforces joint limits and applies temporal smoothing for stability. The output is a full-body pose parameter set θ (SMPL-X compatible) and the head pose \mathbf{T}_H^W used to derive stereo cameras.

Stereo Camera Construction. From the head pose, we construct per-eye camera transforms and transmit them to the GA Backend. We place an interpupillary baseline IPD along the head local +X axis:

$$\mathbf{T}_L^H = \begin{bmatrix} \mathbf{I} & (-\text{IPD}/2, 0, 0)^\top \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad \mathbf{T}_R^H = \begin{bmatrix} \mathbf{I} & (\text{IPD}/2, 0, 0)^\top \\ \mathbf{0}^\top & 1 \end{bmatrix}. \quad (1)$$

World-space eye poses are obtained by composition:

$$\mathbf{T}_e^W = \mathbf{T}_H^W \mathbf{T}_e^H, \quad e \in \{L, R\}. \quad (2)$$

The GA Backend uses the corresponding world-to-camera transforms:

$$\mathbf{T}_{W,e}^C = (\mathbf{T}_e^W)^{-1}. \quad (3)$$

We use the standard pinhole camera model to derive per-eye intrinsics and projection from the headset-provided field of view and render resolution.

Per-Frame Packaging. For each frame, the VR Frontend sends a compact request $\{\text{frame_id}, (\mathbf{T}_{W,e}^C, \mathbf{K}), \theta\}$ to the GA Backend, where frame_id is monotonically increasing, $e \in \{L, R\}$, \mathbf{K} denotes shared camera intrinsics, and θ is the estimated body pose.

Implementation Details. In our reference implementation, the VR Frontend is developed with Unity 6000.1.13f1¹. The IK Module employs Unity Final IK². We use the Meta Quest 3³ as the HMD.

3.2 GA Backend

Role and Interface. The GA Backend receives per-frame requests from the VR Frontend and renders a 3DGS avatar reconstructed from a single image. Each request includes a sequential `frame_id`, per-eye camera parameters for L and R (world-to-camera transforms $\mathbf{T}_{W,e}^C \in \text{SE}(3)$ for $e \in \{L, R\}$ and a shared intrinsic matrix \mathbf{K}), and a full-body pose estimate. The backend returns a stereo pair that is streamed directly to the HMD.

Binocular Batching. We introduce Binocular Batching, which processes the two eye views of a frame within a single batch over a shared 3DGS state. In contrast to a naive stereo pipeline that renders the left and right eye sequentially with duplicated computation, our method stages per-Gaussian attributes $(p_i, \sigma_i, r_i, \rho_i, f_i)$ only once in GPU memory and reuses them across both eyes. Eye-specific projection and visibility $(\mathbf{K}, \mathbf{T}_{W,e}^C)$ are then evaluated separately within the same pass. This reduces redundant memory transfers and per-frame fixed overhead (e.g., kernel launches, data marshalling, cache warm-up), while improving cache coherence and memory locality. As a result, both eye buffers are produced in a single batched pass, yielding a notable reduction in rendering time, which we analyze in detail in subsection 5.2.

Gaussian Avatar Module. The Gaussian Avatar Module estimates canonical 3D Gaussian parameters, SMPL-X rig weights, and animation-ready bindings from a single image. In our implementation, this module is instantiated using the Large Animatable Human Model (LHM) [42] as an off-the-shelf reconstructor. The rest of the GA Backend is agnostic to the reconstructor, and any method that outputs $(\chi, \{w_{ik}\}, \{T_k(\cdot)\})$ can be plugged in.

¹<https://unity.com>

²<https://assetstore.unity.com/packages/tools/animation/final-ik-14290>

³<https://store.facebook.com/quest/products/quest-3>

Tokenization : We encode SMPL-X geometry and image appearance into transformer tokens. Let $X = \{x_j\}_{j=1}^{N_{pts}} \subset \mathbb{R}^3$ be canonical SMPL-X surface samples. Here, N_{pts} denotes the number of sampled surface points, which we set to 2,048 in our implementation. After a sinusoidal positional encoding $\gamma: \mathbb{R}^3 \rightarrow \mathbb{R}^{3L}$ and a modality-specific projection to the channel dimension C , we obtain

$$T^{3D} = \text{MLP}_{\text{proj}}(\gamma(X)) \in \mathbb{R}^{N_{pts} \times C}. \quad (4)$$

From the input image $I \in \mathbb{R}^{H \times W \times 3}$, a frozen body encoder E_{body} produces features in $\mathbb{R}^{N_{\text{body}} \times C_b}$ which are projected to

$$T^{\text{body}} = \text{MLP}_{\text{proj}}(E_{\text{body}}(I)) \in \mathbb{R}^{N_{\text{body}} \times C}. \quad (5)$$

To retain fine facial details, a frozen multi-scale head encoder yields $\{E_{\text{head}}^{(d)}(I) \in \mathbb{R}^{N_{\text{head}}^{(d)} \times C_{h,d}}\}_{d=1}^D$; a fusion module aggregates them into

$$T^{\text{head}} = F_{\text{fusion}}(\{E_{\text{head}}^{(d)}(I)\}_{d=1}^D) \in \mathbb{R}^{N_{\text{head}} \times C}, \quad (6)$$

where N_{head} denotes the post-fusion token count.

Prediction : A Multimodal Body-Head Transformer (MBHT) fuses the three token sets and a decoder predicts canonical Gaussians and rig parameters:

$$(\chi, \{w_{ik}\}, \{T_k(\cdot)\}) = \text{Decoder}(\text{MBHT}(T^{3D}, T^{\text{body}}, T^{\text{head}})). \quad (7)$$

Here, χ contains N Gaussians, $w_{ik} \geq 0$ with $\sum_k w_{ik} = 1$, and $T_k(\theta) \in \text{SE}(3)$ are rigid bone transforms defined in the canonical space.

Concretely, the canonical Gaussian set is

$$\chi = \left\{ (p_i, \sigma_i, r_i, \rho_i, f_i) \right\}_{i=1}^N, \quad (8)$$

where $p_i \in \mathbb{R}^3$ (centroid), $\sigma_i \in \mathbb{R}^3$ (anisotropic scales), $r_i \in \mathbb{S}^3$ (unit quaternion), $\rho_i \in [0, 1]$ (opacity), and $f_i \in \mathbb{R}^C$ (SH appearance). The 3D covariance is

$$\Sigma_i = R(r_i) \text{diag}(\sigma_i^2) R(r_i)^\top, \quad R(r_i) \in \text{SO}(3), \quad (9)$$

with $R(\cdot)$ the quaternion-to-rotation map. To enable animation, the module provides an SMPL-X-compatible rig with blend weights w_{ik} and bone transforms $T_k(\theta) \in \text{SE}(3)$ for pose θ .

Pose conditioning (runtime) : Given the per-frame pose parameters θ from the VR Frontend, we deform the canonical Gaussian set χ into a posed set $\chi' = \text{LBS}(\chi, \theta)$ using linear blend skinning (LBS) with per-Gaussian weights w_{ik} . We define the LBS output for Gaussian i as $(p'_i, \Sigma'_i) = \text{LBS}(p_i, \Sigma_i; \theta)$. Specifically, letting $T_k(\theta) = (R_k(\theta), t_k(\theta))$ denote the pose-dependent transform of bone k and $R_i(\theta)$ the induced blended rotation, we have

$$p'_i = \sum_k w_{ik} (R_k(\theta) p_i + t_k(\theta)), \quad (10)$$

$$\Sigma'_i = R_i(\theta) \Sigma_i R_i(\theta)^\top. \quad (11)$$

Splatting and compositing : We render posed Gaussians using the standard 3D Gaussian Splatting rasterization [21].

$$\mathbf{I}_e = \text{Render}(\text{LBS}(\chi, \theta), \mathbf{K}, \mathbf{T}_{W,e}^C), \quad e \in \{L, R\}, \quad (12)$$

where $\text{LBS}(\chi, \theta)$ denotes the posed Gaussian set after the LBS step above, and $(\mathbf{K}, \mathbf{T}_{W,e}^C)$ are the per-eye camera parameters. Under Binocular Batching, per-Gaussian attributes are staged once and reused for both L and R within a single batch, while per-eye projection and view-dependent shading are evaluated per eye in the same pass.

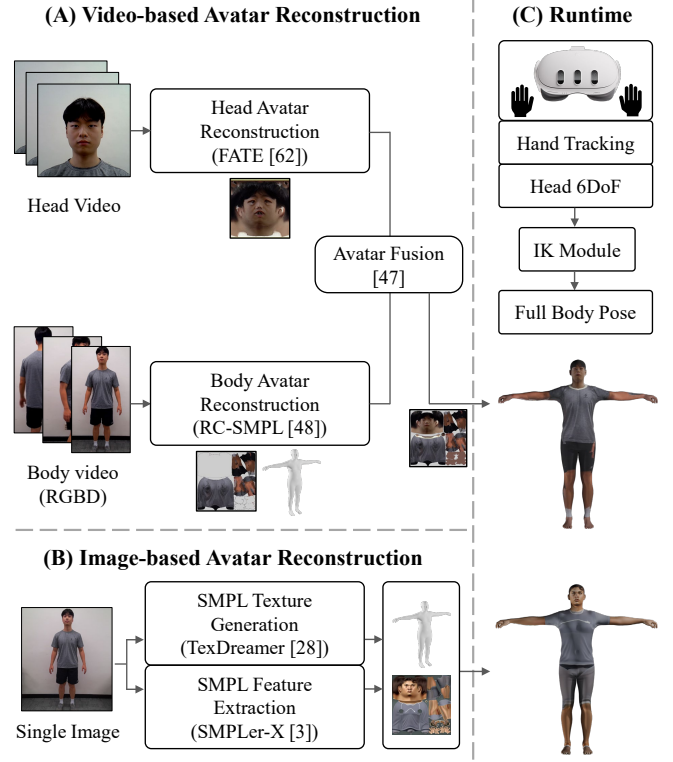


Fig. 3: **Baseline conditions.** (A) *video-based reconstruction*: Implemented following prior works [47, 48, 62]. (B) *image-based reconstruction*: Implemented following prior works [3, 28]. (C) *Runtime process*: The IK Module estimates a full-body pose from head 6DoF and hand-tracking signals to animate the avatar in real time.

Encoding and Delivery to HMD. Rendered frames are converted to 8-bit sRGB and JPEG-compressed at a fixed quality; the left/right payloads, together with `frame_id`, are transmitted in a single response and delivered directly to the HMD. The stateless per-frame design admits deterministic dropping of late frames on the receiver.

The GA Backend renders the 3DGS avatar with an alpha channel to support optional avatar compositing when a 3DGS scene/background is included. In our study, we disable background compositing and use a black background to minimize background influence and focus the evaluation on the avatar itself; accordingly, we stream RGB-only JPEG frames to reduce bandwidth.

Algorithmic Outline.

1. Receive $\{\text{frame_id}, (\mathbf{T}_{W,e}^C, \mathbf{K}), \theta\}$ from the VR Frontend.
2. Construct a single-batch stereo rendering context and apply pose conditioning to the canonical 3DGS avatar, deforming Gaussian primitives according to θ .
3. Execute one batched splatting pass with *Binocular Batching* to simultaneously render left and right eye images $[L, R]$.
4. JPEG-compress the two rendered eye buffers and package them with `frame_id`.
5. Stream the compressed response directly back to the HMD for display.

Implementation Details. GA Backend is implemented in Python 3.9⁴, PyTorch 2.3.0⁵, and PyTorch3D⁶. Rendering runs on a single NVIDIA RTX 4090 GPU in real time.

⁴<https://www.python.org/>

⁵<https://pytorch.org/>

⁶<https://pytorch3d.org/>



Fig. 4: (A) User study setup with participants performing poses from pre-recorded voice instructions. (B) Avatar type factor with two levels: *self* (left, resembling the participant) and *other* (right, resembling another person).

4 EXPERIMENTS

4.1 Baselines

To evaluate our system, we compare against mesh-based full-body avatar reconstruction methods using the SMPL-X [38] model. Figure 3 illustrates the baseline systems constructed for comparison: (A) video-based mesh avatar reconstruction, and (B) image-based mesh avatar reconstruction. For 3DGS avatar generation in our system, VRGaussianAvatar (*Ours*), a frontal image of each subject in an A-pose with arms slightly extended was captured. This single image served as the input for generating a full-body 3DGS avatar.

In the case of video-based mesh avatar reconstruction (*video*), we separately utilized head and body videos. For the head, we recorded a one-minute video of the Facial Action Coding System (FACS) [8]-based facial expressions and reconstructed it with FATE [62] to obtain a textured FLAME [26] model. For the body, we captured RGBD videos using the Azure Kinect DK⁷ and reconstructed the textured SMPL-X model with RC-SMPL [48]. The head and body textures were subsequently merged using the method of [47].

For image-based mesh avatar reconstruction (*image*), we used the same single input image employed in VRGaussianAvatar. We adopted the off-the-shelf SMPL-X inference model SMPLer-X [3] for mesh recovery, and generated textures with TexDremer [28], an image-based SMPL-X texture generator.

Experimental Details. All baselines and the VRGaussianAvatar runtime were executed on a desktop PC equipped with an Intel Core i9-14900K CPU, an NVIDIA RTX 4090 GPU, and 64 GB of RAM. For avatar reconstruction, both the image-based mesh baseline and our system’s Gaussian Avatar Module were run on the same desktop PC as the runtime environment. The image-based baseline required approximately 30 seconds for SMPL-X parameter estimation and 19 seconds for texture generation. Our Gaussian Avatar Module reconstructed a full-body 3D Gaussian Avatar within 10 seconds. For the video-based mesh baseline, body reconstruction with RC-SMPL took about 30 seconds, and avatar integration required an additional 2 seconds. Facial reconstruction with FATE was performed on a separate workstation equipped with an NVIDIA A40 GPU and required about 12 hours.

4.2 User Study

We conducted a human-subject study to evaluate whether our system can effectively generate realistic user avatars for VR. The evaluation considered two perspectives: (1) whether a self-resembling avatar is perceived as one’s own, and (2) whether an avatar resembling another person can be accepted as one’s own. Based on these perspectives, we compared our system with two baseline methods.

Study Design and Task.

The experimental factors and conditions were organized as follows. The first factor was avatar type with two levels: (1) self-representative avatar (*self*) and (2) other-representative avatar (*other*). For the *other* avatar type, we used avatars of non-participant models matched in gender, ethnicity, and similar age to the participants (Figure 4 (B)). The second factor was avatar reconstruction method with three conditions related to the generation process: (1) *Ours* (VRGaussianAvatar), (2) *video* (video-based avatar reconstruction) [47, 48, 62], and (3) *image* (image-based avatar reconstruction) [3, 28] methods. Both of the avatar type and method factors were within-subject factors, our experiment employed a 2×3 within-subject factorial design. To compare method conditions under the self and other perspectives, avatar type was counterbalanced across participants, and the three method conditions within each type were also counterbalanced using a Latin Square design.

To evaluate the appearance and movements of avatars in VR, we adapted a commonly used avatar control task from prior studies assessing virtual embodiment [7, 44, 54]. In this task, participants wearing an HMD experienced their avatars in mirror mode and were instructed through audio cues to make specific body poses, while observing how these poses were reflected in their avatar [44]. We chose the mirror-based observation (rather than a purely first-person view) because our primary goal was to let participants directly inspect the reconstructed avatar’s overall appearance and motion quality as a self-representation. Building on the previous work [7], which extended the original task [44] by adding side-view observation instructions, we designed the task to better fit our study context and to ensure that participants could sufficiently observe the avatars. For instance, participants were encouraged to freely explore their avatars (e.g., “Move your arms and take any posture you want.”) and were also given additional instructions to observe their avatars not only from the front, but also from the left, right, and rear views.

Dependent Variables. In this study, the Sense of Embodiment (SoE) was measured as the primary dependent variable, as the study aimed to evaluate whether our realistic avatars’ appearance and movements conveyed appropriate body representation and perception [2, 44]. Based on the previous studies that explored overall perception and embodiment of avatars in VR [30, 31, 33, 54, 58], we adopted two representative measurements: (1) the Virtual Embodiment Questionnaire (VEQ) by Roth and Latoschik [44] and (2) VEQ+ by Fiedler et al. [11]. The VEQ, which is specifically designed to assess SoE in VR environments, consists of three latent factors. First, Virtual Body Ownership (VBO) refers to the perception of a virtual body as one’s own. Second, Agency (AG) denotes the perception of control over the virtual body, and the last Change (CH) captures alterations in the perceived body schema. The VEQ+ extends this framework by incorporating two additional dimensions: Self-location (the feeling of being located at the position of one’s body) and Self-identification (related to personalization and resemblance). Under these dimensions, VEQ+ has three subscales—Self-Location (SL), Self-Attribution (SA), and Self-Similarity (SS). Both the VEQ and VEQ+ comprise 12 items, for a total of 24 items.

The Virtual Human Plausibility Questionnaire (VHPQ) measures the extent to which a virtual avatar’s appearance and behavior are perceived as plausible, coherent, and consistent with the VR environment [29]. Considering the objective and setup of our system evaluation, we only used the Appearance and Behavior Plausibility (ABP) subscale of the VHPQ, which consisted of six items. In addition, following earlier research [44], we also included three self-reported items to more di-

⁷<https://learn.microsoft.com/en-us/azure/kinect-dk/>

rectly compare the realism, naturalness, and visuomotor synchrony of avatar motion behavior: “*The virtual avatar’s movements were realistic*,” “*The virtual avatar’s movements were naturalistic*,” and “*The virtual avatar’s movements were in synchrony with my own movements*.” All questionnaire measures were assessed on a 7-point Likert scale (1: Strongly disagree – 7: Strongly agree). For the post-experiment interview, we designed open-ended questions with reference to Menzel et al. [33] to collect qualitative feedback on avatar fidelity, suitability, and realism when observing both *self* and *other* avatar types. Participants were also asked about their subjective opinions on differences between avatars and their preferences.

Procedure. This study was approved by the Institutional Review Board (IRB). Each participant attended two sessions. The first session was for data acquisition for the *video* method, where participants’ facial data were scanned (pre-processing required over 12 hours); They also received an explanation of the study, provided informed consent, and completed a demographic questionnaire. The second session, conducted at least one day later, was the main experiment. At the beginning, participants’ bodies were scanned to generate avatars for the *Ours* and *image* conditions, while the *video* avatars were generated by combining the pre-processed facial model with the newly scanned body.

After avatar generation, participants wore an HMD and performed the avatar control task while standing in an empty experimental space (Fig. 4 (A)). Following pre-recorded voice instructions, they assumed various body poses and observed their virtual avatars mirroring their poses in real-time. After each method condition, they removed the HMD and completed the post-task questionnaire. Participants first completed the three method conditions (*Ours*, *video*, *image*) arranged in a balanced order under one avatar type, and then repeated the sequence under the other type, resulting in six evaluations in total. Finally, a post-experiment interview was conducted to collect qualitative feedback on their experiences with each method under both *self* and *other* conditions.

Participants. We recruited 24 participants (12 male, 12 female) through the campus website. Their ages ranged from 22 to 34 years ($M = 27.67$, $SD = 3.17$). Regarding prior experience with avatar-mediated applications (e.g., social networking services, games), 17 participants reported using such applications fewer than five times (70.83%), whereas seven participants reported using them five times or more (29.17%). In contrast, most participants (20, 83.3%) indicated that they had used AR/VR technologies and wearable devices more than five times, while only four participants reported fewer than five times of AR/VR experience (16.67%).

4.3 Quantitative Experiments

We conducted a quantitative evaluation of the Gaussian Avatar Module, a core component of VRGaussianAvatar responsible for avatar representation. For comparison, we employed video-based [47, 48, 62] and image-based [3, 28] avatar reconstruction methods that are also applicable in VR, consistent with our user study setup. We also measured the rendering speed at different resolutions to assess the contribution of Binocular Batching for real-time rendering.

Datasets. We recruited 24 participants (12 male and 12 female) for evaluation. For both VRGaussianAvatar and the image-based baseline, a single source image was captured with the subject standing in an A-pose (arms lowered and slightly apart from the body). For the video-based baseline, we recorded videos for facial and body reconstruction. Facial reconstruction used a one-minute RGB video containing diverse expressions based on the FACS [8], while body reconstruction relied on a 30-second RGBD video capturing both the frontal and rear views of the subject. For fair evaluation, rendered results from each method were cropped to the full-body and compared against the corresponding ground-truth images.

Metrics. We employed four widely used metrics for evaluating avatar reconstruction quality: L1 distance, LPIPS [63], PSNR [15], and

SSIM [55]. L1 distance measures pixel-wise absolute error between the rendered image and ground truth. LPIPS (Learned Perceptual Image Patch Similarity) assesses perceptual similarity based on deep neural network features, better reflecting human visual perception. PSNR (Peak Signal-to-Noise Ratio) quantifies reconstruction quality in terms of signal fidelity, with higher values indicating less distortion. SSIM (Structural Similarity Index) evaluates structural consistency and perceptual quality, particularly in terms of luminance, contrast and structure.

5 RESULTS

5.1 User Study Analysis

For subjective measures, a two-way repeated measures ANOVA with the Aligned Rank Transform (ART) was used for non-parametric factorial analysis ($\alpha = .05$) [57]. The ART method has been widely utilized in human-subject research analysis because it is specialized in multi-factorial non-parametric analysis and accommodates ordinal and continuous variables [57]. All post-hoc comparisons were conducted using the ART-C procedure [9] with Bonferroni corrected p-values. The internal consistency among Likert items was examined by the Cronbach’s alpha reliability coefficient.

5.1.1 Virtual Embodiment

VEQ. The results for the VEQ and its subscales are shown in Fig. 6. The VEQ score, which aggregates all three subscales, showed a reliable consistency ($\alpha_{VEQ} = .825$), and the three subscales also showed an acceptable value ($\alpha_{VBO} = .861$; $\alpha_{AG} = .837$; $\alpha_{CH} = .890$). We found a main effect of avatar type ($F_{1,23} = 8.464$, $p = .008$, $\eta_p^2 = .269$), with *self* rated higher than *other* avatar types. There were no significant main effect of method ($F_{2,46} = 1.082$, $p = .348$, $\eta_p^2 = .045$), nor interaction effect between the two factors ($F_{2,46} = 2.493$, $p = .094$, $\eta_p^2 = .098$).

For VBO, significant main effects were found for avatar type ($F_{1,23} = 15.064$, $p < .001$, $\eta_p^2 = .396$) and method ($F_{2,46} = 21.903$, $p < .001$, $\eta_p^2 = .488$). Post-hoc tests indicated that *self* avatars were rated higher than *other*, and *Ours* method induced higher ownership than both *video* and *image* (all $p < .001$), while *video* and *image* did not differ. No interaction effect was observed ($F_{2,46} = 1.419$, $p = .252$). AG showed significant main effects of avatar type ($F_{1,23} = 4.323$, $p = .049$, $\eta_p^2 = .158$) and method ($F_{2,46} = 6.041$, $p = .005$, $\eta_p^2 = .208$). Post-hoc comparisons revealed that *self* avatars scored higher than *other* avatars, and that *Ours* conveyed higher agency than both *video* ($p = .031$) and *image* ($p = .006$), while *video* and *image* were not different. The interaction was not significant ($F_{2,46} = 0.841$, $p = .438$). The CH score, with lower values reflecting less perceived change from one’s own body and thus a better representation of the actual user, showed a significant main effect of method ($F_{2,46} = 13.728$, $p < .001$, $\eta_p^2 = .374$), but no main effect of avatar type. Pairwise comparisons showed that *Ours* yielded lower CH scores than both *video* ($p = .001$) and *image* ($p < .001$), whereas *video* and *image* did not differ. In addition, a significant interaction effect was observed between two factors ($F_{2,46} = 7.932$, $p = .001$, $\eta_p^2 = .256$). Follow-up analyses showed that in the *self* condition, *Ours* produced lower CH scores than both *video* and *image* (all $p < .001$), while in the *other* condition, *Ours* was only lower than *image* ($p = .024$).

VEQ+. VEQ+ ($\alpha_{VEQ+} = .937$) and its subscales all showed an acceptable Cronbach’s alpha ($\alpha_{SL} = .765$; $\alpha_{SA} = .889$; $\alpha_{SS} = .965$). The results of VEQ+ are also in Fig. 6. The overall VEQ+ score revealed significant main effects of avatar type ($F_{1,23} = 57.715$, $p < .001$, $\eta_p^2 = .715$) and method ($F_{2,46} = 49.487$, $p < .001$, $\eta_p^2 = .683$), and a significant interaction effect ($F_{2,46} = 16.155$, $p < .001$, $\eta_p^2 = .413$). Post-hoc tests showed that *self* avatars scored higher than *other*, and *Ours* was significantly higher than both *video* and *image* (all $p < .001$). Also, *video* was rated higher than *image* ($p < .001$). Within the *self* condition, *Ours* was significantly higher than both *video* and *image* (all $p < .001$), and *video* was rated higher than *image* ($p = .005$). Within the *other* condition, *Ours* again scored significantly higher than

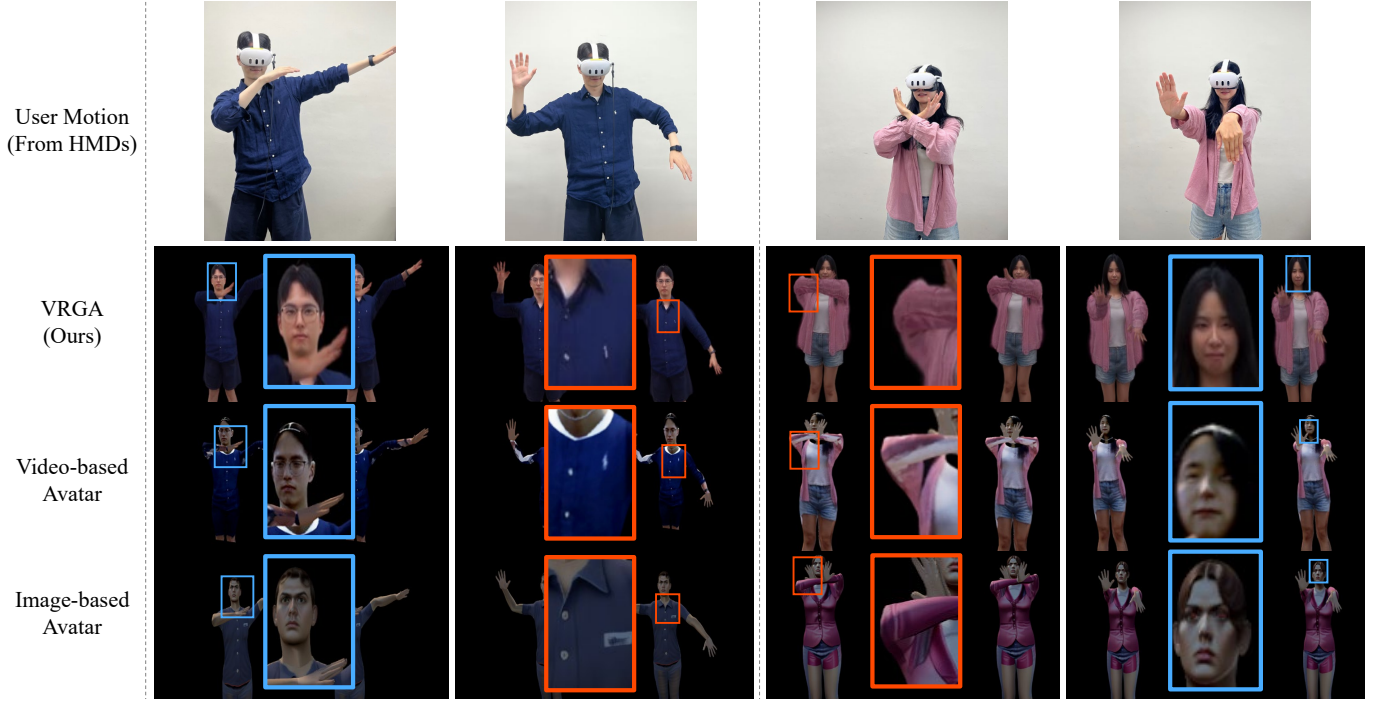


Fig. 5: Qualitative comparison of VRGaussianAvatar with baseline avatar representations under identical HMD-only control. Compared to baselines, VRGaussianAvatar typically preserves sharper appearance details in close-up regions, including head/face (sky-blue boxes) and clothing (orange boxes). All methods are rendered stereoscopically in real time, enabling interactive avatar control in VR.

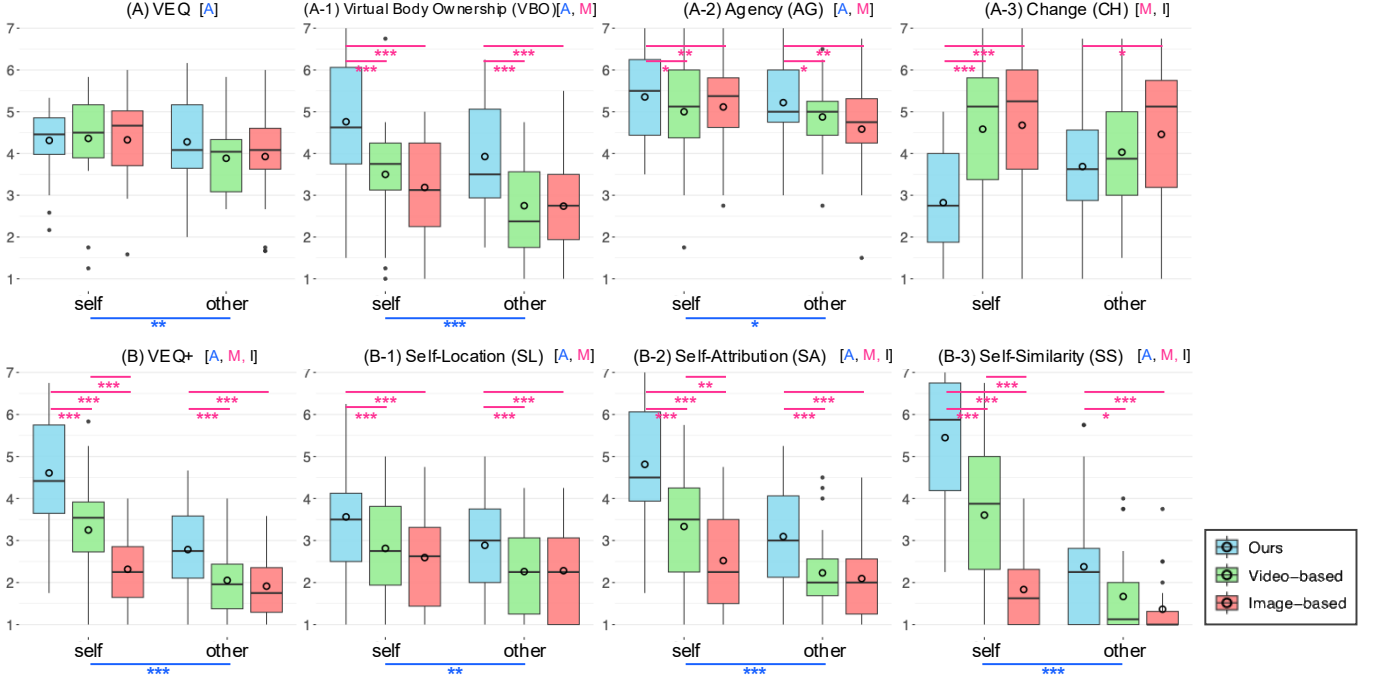


Fig. 6: Results for (A)–(A-3) Virtual Embodiment Questionnaire and its subscales; (B)–(B-3) VEQ+ and its subscales. (A and M: significant main effect of avatar type and method, respectively; I: significant interaction effect between avatar type and method)

video and image ($p < .001$), while the difference between video and image was not significant.

SL showed significant main effects of avatar type ($F_{1,23} = 9.389$, $p = .005$, $\eta_p^2 = .290$) and method ($F_{2,46} = 12.948$, $p < .001$, $\eta_p^2 = .360$). Post-hoc comparisons indicated that *self* avatars were rated higher than *other* avatars, and *Ours* showed significantly higher SL

than *video* ($p = .001$) and *image* ($p < .001$), whereas *video* and *image* did not differ. No interaction effect was found. SA revealed significant main effects of avatar type ($F_{1,23} = 45.904$, $p < .001$, $\eta_p^2 = .666$) and method ($F_{2,46} = 44.525$, $p < .001$, $\eta_p^2 = .659$), as well as a significant interaction effect ($F_{2,46} = 10.653$, $p < .001$, $\eta_p^2 = .317$). Post-hoc analysis showed that *self* type was higher than *other* type; *Ours* method

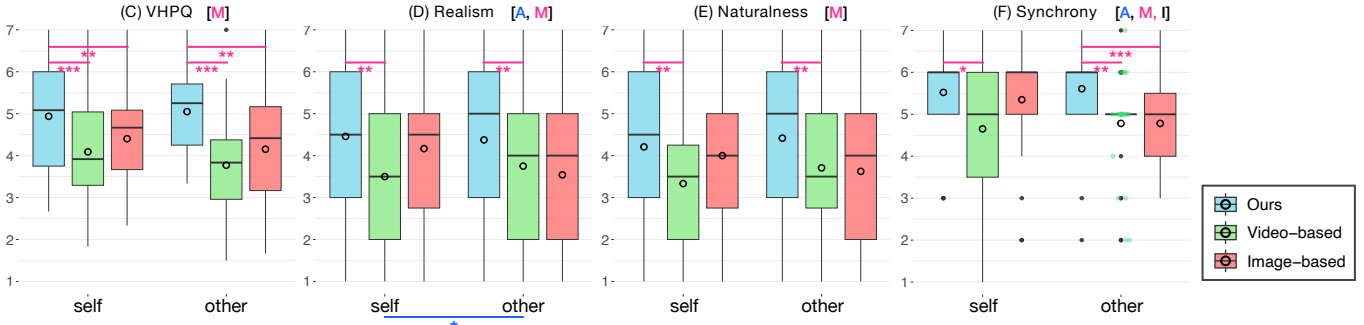


Fig. 7: Results for (C) Virtual Human Plausibility Questionnaire; (D)–(F) Additional subjective ratings. In (F), green jittered points indicate individual responses for the video-based method in the other avatar type, shown to complement the box plot where the interquartile range collapsed to a single value. (A and M: significant main effect of avatar type and method, respectively; I: significant interaction effect between avatar type and method)

was rated significantly higher than *video* and *image* (all $p < .001$), and *video* was also higher than *image* ($p = .006$). In both the *self* and *other* conditions, *Ours* induced significantly higher scores than *video* and *image* (all $p \leq .001$), and *video* was higher than *image* but only in the *self* condition ($p = .005$). Similarly, analysis of SS revealed significant main effects of avatar type ($F_{1,23} = 71.296$, $p < .001$, $\eta_p^2 = .756$), method ($F_{2,46} = 63.800$, $p < .001$, $\eta_p^2 = .735$), and their interaction ($F_{2,46} = 25.390$, $p < .001$, $\eta_p^2 = .525$). Post-hoc comparisons first indicated that *self* type was higher than *other* type. Moreover, *Ours* had significantly higher SS than *video* and *image*, and *video* was also higher than *image* (all $p < .001$). Within the *self* type, *Ours* perceived more similar with the participants than both *video* and *image*, and *video* was rated higher than *image* (all $p < .001$). In the *other* type, *Ours* also scored higher similarity than *video* ($p = .027$) and *image* ($p < .001$), while *video* and *image* had no difference.

5.1.2 Virtual Avatar Plausibility and Motion Behavior

VHPQ. The results regarding the VHPQ and self-reported measures are in Fig. 7. With respect to the ABP in VHPQ, it showed a reliable Cronbach’s alpha ($\alpha_{VHPQ} = .900$). A main effect of method was observed ($F_{2,46} = 13.126$, $p < .001$, $\eta_p^2 = .363$), with *Ours* induced higher appearance and behavior plausibility than *video* ($p = .001$) and *image* ($p = .005$). No effects of avatar type or interaction were found.

Self-reported Motion Behavior. For motion realism, significant main effects of avatar type ($F_{1,23} = 4.656$, $p = .042$, $\eta_p^2 = .168$) and method ($F_{2,46} = 5.244$, $p = .009$, $\eta_p^2 = .186$) were observed, while the interaction effect was not significant. The pairwise post-hoc tests indicated that *self* avatars were higher than *other* avatars, and *Ours* induced higher behavior realism than *video* ($p = .009$). No other differences were significant. Regarding naturalness, a significant main effect of method was only found ($F_{2,46} = 4.950$, $p = .011$, $\eta_p^2 = .177$). Similar with the realism, post-hoc analysis revealed that *Ours* had significantly higher naturalness than *video* ($p = .009$), while *image* did not differ from either. No effect of avatar type or interaction was found. Lastly, one participant’s data contained an error and was excluded from the synchrony analysis. Synchrony showed a main effect of method ($F_{2,44} = 10.604$, $p < .001$, $\eta_p^2 = .325$), and an interaction with avatar type ($F_{2,44} = 3.459$, $p = .040$, $\eta_p^2 = .136$). In post-hoc analysis, it was revealed that *Ours* was rated as having higher motion synchrony than both *video* and *image* (all $p < .001$), and also *video* was higher than *image* ($p = .012$). In the *self* condition, *Ours* method had higher values than *video* ($p = .031$). In the *other* condition, *Ours* was higher than both *video* ($p = .002$) and *image* ($p < .001$), while other pairs were not different.

5.2 Quantitative Results

Table 1 summarizes the quantitative evaluation across different conditions. Our method (VRGA) achieves the best performance in all

Table 1: Quantitative comparison of rendering quality across different conditions. Video-based avatars are constructed following [47, 48, 62], and image-based avatars are constructed following [3, 28].

Condition	L1 ↓	LPiPS ↓	PSNR ↑	SSIM ↑
Video-based Avatar	0.042	0.099	16.33	0.927
Image-based Avatar	0.043	0.103	16.74	0.917
VRGA (Ours)	0.033	0.082	17.92	0.934

Table 2: Rendering performance across different resolutions with and without Binocular Batching (BB). Binocular Batching enables efficient real-time stereoscopic rendering. Resolution indicates the per-eye resolution, and FPS is measured under binocular rendering.

Resolution	Binocular Batching	LPiPS ↓	Render Time	FPS ↑
512×512	w/ BB	0.085	21.78 ms	45.91
	w/o BB		37.47 ms	26.69
768×768	w/ BB	0.084	23.19 ms	43.12
	w/o BB		38.33 ms	26.09
1024×1024	w/ BB	0.082	25.45 ms	39.29
	w/o BB		38.80 ms	25.77

metrics, demonstrating consistent advantages over both baselines. Between the two baselines, the video-based avatar shows better results than the image-based avatar in terms of L1, LPiPS, and SSIM.

Table 2 presents image similarity, rendering time, and FPS across different resolutions with and without Binocular Batching (BB). By exploiting the fact that stereoscopic rendering in VR shares SMPL-X parameters and dynamic 3DGS attributes, our system renders both camera views simultaneously with a batch size of two, enabling high-quality avatar rendering in real time. For the user study, we employed Binocular Batching at a resolution of 1024×1024, which provided real-time performance and visual quality.

Figure 5 provides representative qualitative examples that visually corroborate these quantitative improvements, showing clearer appearance details for VRGaussianAvatar under identical HMD-only control. More qualitative results are provided in Appendix B.

6 DISCUSSION

The goal of this work is to enable practical, high-fidelity avatar representation for VR applications by addressing the limitations of existing video- and image-based approaches. To this end, we introduce VRGaussianAvatar, an HMD-only controllable avatar pipeline that integrates 3D Gaussian Splatting into an interactive VR system, and evaluate whether

these technical contributions translate into perceptually meaningful improvements through a user study.

Our method produced stronger virtual embodiment than video and image-based approaches, particularly in body ownership and agency. Participants reported that the avatar looked and moved like their own body, with less alteration and greater resemblance. While aggregated VEQ scores did not differ significantly, subscales revealed that the sense of “my body” was more clearly enhanced with our method (“*felt the closest to how I actually move*,” P7). Further analyses of VEQ+ showed notable advantages in self-location and self-identification (self-attribution, and self-similarity). Participants described feeling that their body shifted into the avatar’s position in VR and that the avatar both resembled and responded to them. Even in other-representative conditions, where similarity scores were relatively low, our method still surpassed alternatives by producing more human-like and relatable avatars. By contrast, video-based avatars were described as “*overall texture seemed reflected but with unnatural movement*,” and image-based as “*only reflecting clothing color, not my actual features*” (P2).

Beyond embodiment, our system achieved the highest plausibility and realism, with more natural and synchronized movements than video-based avatars. Although some participants acknowledged that image-based avatars behaved smoothly or without latency (P13, P18), they emphasized that our method offered more consistent control and stronger human-likeness (“*moved naturally and felt alive*,” P15). Finally, these advantages translated into higher preference and practical usability. Most participants (79.2%) selected our method as the avatar they would most prefer and want to use in a virtual space, describing it as “*resembled me the most and felt natural*” (P20, P23). Even when evaluating “other” avatars, Ours was the most appealing to participants (50%) while maintaining higher plausibility and synchrony, underscoring its applicability for social VR scenarios. Taken together, these findings show that our method better represents virtual embodiment, identification, and realism than baseline approaches, providing a solid foundation for both self and other-representative avatars in VR.

These results indicate that the perceptual advantages observed in the user study may be associated with the technical design of the proposed pipeline. By separating real-time pose estimation from high-cost Gaussian rendering through the VR Frontend and GA Backend architecture, the system maintains stable and responsive control while preserving visual fidelity, which may help support users’ sense of embodiment and agency. In addition, Binocular Batching enables consistent stereoscopic rendering without redundant computation, allowing 3DGS-based avatars to remain interactive and perceptually plausible. As such, VRGaussianAvatar provides a practical foundation for deploying high-fidelity, controllable avatars in scenarios where both visual realism and embodied interaction play central roles; it is expected to show particularly high applicability in multi-user immersive environments in which users must continuously perceive both themselves and others as coherent agents (e.g., social VR, remote collaboration, or games). While further validation with larger samples and extended tasks is needed, the ability to support both self- and other-representative avatars within a single, efficient pipeline has the potential to broaden its applicability.

6.1 Limitations and Future Work

Our system has several limitations. First, because we do not integrate facial sensing signals, the avatar does not support dynamic facial expressions, gaze, or blinking. Second, our reconstruction assumes a controlled input image that is favorable for one-shot full-body reconstruction; performance may degrade under substantial occlusions or poses that deviate from the capture guideline. Third, while GPU-side rendering runs in real time on a desktop GPU, the end-to-end system latency is subject to streaming overhead. While our implementation runs on localhost, deploying the GA Backend over a remote network would increase end-to-end latency and may require bitrate–quality trade-offs (e.g., lower resolution or stronger compression) to maintain interactivity. Finally, our user study primarily relies on mirror-based observation and therefore does not directly evaluate embodiment outcomes from an egocentric (first-person) viewpoint.

Future work will aim to improve practicality and generalizability by addressing the limitations above. First, for social VR scenarios, we plan to support dynamic facial control by incorporating head-avatar reconstruction/control techniques (e.g., VOODOO XP [51]). Second, to reduce reliance on controlled capture, we will explore robustness strategies for more unconstrained inputs. Third, to move toward standalone deployment, we will investigate on-device avatar representation by adopting lightweight 3DGS avatar techniques (e.g., SqueezeMe [16]). Fourth, we will conduct additional user studies that directly evaluate embodiment from a first-person viewpoint.

6.2 Ethical Considerations

Because our system can generate high-fidelity avatars from a single image, it may be misused for non-consensual digital cloning or impersonation. To mitigate such risks, we restrict the integrated system to research and educational use and require written consent from the subject prior to avatar creation, including a clear specification of the intended use and duration. In addition, XR deployments that present reconstructed avatars should explicitly disclose the driving user to reduce deception risks. Our user study was conducted under Institutional Review Board (IRB) approval and followed standard consent and privacy procedures. If considering broader deployment, safeguards will likely require a combination of technical and procedural measures, such as consent protocols, identity verification, and standardized watermarking.

7 CONCLUSION

Building on these findings, we presented a system for real-time full-body 3DGS avatar representation in VR environments that can be controlled solely through internal sensors of commercial HMDs. Our framework integrates a VR Frontend, which estimates controllable full-body motion from these sensor signals, with a GA Backend that performs efficient stereoscopic rendering via Binocular Batching. This parallel design enables photorealistic and responsive avatar experiences without the need for additional external trackers, lowering the barrier for practical VR deployment. In addition, our quantitative evaluations and user studies confirm that VRGaussianAvatar not only achieves more photorealistic avatar representation compared to mesh-based avatar systems reconstructed from images and videos, but also provides higher levels of virtual embodiment and plausibility for users.

In summary, VRGaussianAvatar demonstrates that 3D Gaussian Splatting can be effectively extended to real-time, photorealistic, and HMD-only controllable avatar representation in VR. We believe this work highlights the feasibility of real-time 3DGS-based avatars and opens the door to richer and more engaging social interaction in immersive environments.

ACKNOWLEDGMENTS

This work was supported by the National Research Council of Science Technology (NST) grant by the Korea government (MSIT) (No. CRC21015). This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program (IITP-2022(2025)-RS-2022-00156435) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was supported by the Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-25441313, Professional AI Talent Development Program for Multimodal AI Agents). This research was financially supported by Hansung University for Hyunjeong Kim. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).

REFERENCES

- [1] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8387–8397, 2018. 1, 2

- [2] S. Benford, J. Bowers, L. E. Fahlén, C. Greenhalgh, and D. Snowden. User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 242–249, 1995. 5
- [3] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023. 1, 4, 5, 6, 8
- [4] A. Chatziagapi, B. Chaudhuri, A. Kumar, R. Ranjan, D. Samaras, and N. Sarafianos. Talkinnerf: Animatable neural fields for full-body talking humans. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024. 1, 2
- [5] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. *arXiv preprint arXiv:2312.04558*, 2023. 2
- [6] S. Cho, S.-w. Kim, J. Lee, J. Ahn, and J. Han. Effects of volumetric capture avatars on social presence in immersive virtual environments. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 26–34. IEEE, 2020. 2
- [7] T. D. Do, C. I. Protko, and R. P. McMahan. Stepping into the right shoes: The effects of user-matched avatar ethnicity and gender on sense of embodiment in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2434–2443, 2024. 5
- [8] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 5, 6
- [9] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*, pp. 754–768, 2021. 6
- [10] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1
- [11] M. L. Fiedler, E. Wolf, N. Döllinger, M. Botsch, M. E. Latoschik, and C. Wienrich. Embodiment and personalization for self-identification with virtual humans. In *2023 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pp. 799–800. IEEE, 2023. 5
- [12] X. Gao, C. Zhong, J. Xiang, Y. Hong, Y. Guo, and J. Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 1
- [13] Y. He, X. Gu, X. Ye, C. Xu, Z. Zhao, Y. Dong, W. Yuan, Z. Dong, and L. Bo. Lam: Large avatar model for one-shot animatable gaussian head. *arXiv preprint arXiv:2502.17796*, 2025. 1
- [14] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9352–9364, 2023. 2
- [15] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6
- [16] F. Iandola, S. Pidhorskyi, I. Santesteban, D. Gupta, A. Pahuja, N. Bartolovic, F. Yu, E. Garbin, T. Simon, and S. Saito. Squeezeeme: Mobile-ready distillation of gaussian full-body avatars. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers*, pp. 1–11, 2025. 2, 9
- [17] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *arXiv preprint arXiv:2212.10550*, 2022. 1, 2
- [18] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–1, 2024. 1, 2
- [19] J.-N. Kaiser, S. Kimmel, E. Licht, E. Landwehr, F. Hemmert, and W. Heuten. Get real with me: Effects of avatar realism on social presence and comfort in augmented reality remote collaboration and self-disclosure. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2025. 2
- [20] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1
- [21] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1, 4
- [22] H. Kim and I.-K. Lee. Is 3dgs useful?: Comparing the effectiveness of recent reconstruction methods in vr. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 71–80. IEEE, 2024. 1
- [23] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020. 1
- [24] Y.-C. Lee, Y.-T. Chen, A. Wang, T.-H. Liao, B. Y. Feng, and J.-B. Huang. Vividdream: Generating 3d scene with ambient dynamics. *arXiv preprint arXiv:2405.20334*, 2024. 1
- [25] K. Li, T. Rolff, S. Schmidt, R. Bacher, S. Frintrop, W. Leemans, and F. Steinicke. Immersive neural graphics primitives. *arXiv preprint arXiv:2211.13494*, 2022. 2
- [26] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 5
- [27] X. Liu, C. Wu, X. Liu, J. Liu, J. Wu, C. Zhao, H. Feng, E. Ding, and J. Wang. Gea: Reconstructing expressive 3d gaussian avatar from monocular video. *arXiv preprint arXiv:2402.16607*, 3, 2024. 2
- [28] Y. Liu, J. Zhu, J. Tang, S. Zhang, J. Zhang, W. Cao, C. Wang, Y. Wu, and D. Huang. Texdreamer: Towards zero-shot high-fidelity 3d human texture generation. In *European conference on computer vision*, pp. 184–202. Springer, 2024. 2, 4, 5, 6, 8
- [29] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik. Virtual human coherence and plausibility—towards a validated scale. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 788–789. IEEE, 2022. 5
- [30] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik. From 2d-screens to vr: Exploring the effect of immersion on the plausibility of virtual humans. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2024. 5
- [31] D. Mal, E. Wolf, N. Döllinger, C. Wienrich, and M. E. Latoschik. The impact of avatar and environment congruence on plausibility, embodiment, presence, and the proteus effect in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2358–2368, 2023. 5
- [32] Y. Men, B. Lei, Y. Yao, M. Cui, Z. Lian, and X. Xie. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9981–9991, June 2024. 2
- [33] T. Menzel, E. Wolf, S. Wenninger, N. Spinczyk, L. Holderrieth, C. Wienrich, U. Schwanecke, M. E. Latoschik, and M. Botsch. Avatars for the masses: smartphone-based reconstruction of humans for virtual reality. *Frontiers in Virtual Reality*, 6:1583474, 2025. 5, 6
- [34] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2
- [35] G. Moon, T. Shiratori, and S. Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024. 1, 2
- [36] A. Moreau, J. Song, H. Dhamo, R. Shaw, Y. Zhou, and E. Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 788–798, 2024. 2
- [37] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2
- [38] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. 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. 1, 2, 5
- [39] S. Peng, W. Xie, Z. Wang, X. Guo, Z. Chen, B. Yang, and X. Dong. Rmavatar: Photorealistic human avatar reconstruction from monocular video based on rectified mesh-embedded gaussians. *Graphical Models*, 139:101266, 2025. 2
- [40] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *arXiv preprint arXiv:2312.02069*, 2023. 2
- [41] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024. 2
- [42] L. Qiu, X. Gu, P. Li, Q. Zuo, W. Shen, J. Zhang, K. Qiu, W. Yuan, G. Chen, Z. Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 1, 2, 3, 12
- [43] L. Qiu, P. Li, Q. Zuo, X. Gu, Y. Dong, W. Yuan, S. Zhu, X. Han, G. Chen, and Z. Dong. Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images. *arXiv preprint arXiv:2506.13766*, 2025. 2

- [44] D. Roth and M. E. Latoschik. Construction of the virtual embodiment questionnaire (veq). *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3546–3556, 2020. 5
- [45] G. Sim and G. Moon. Persona: Personalized whole-body 3d avatar with pose-driven deformations from a single image. *arXiv preprint arXiv:2508.09973*, 2025. 1, 2
- [46] H. Song. Toward realistic 3d avatar generation with dynamic 3d gaussian splatting for ar/vr communication. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 869–870. IEEE, 2024. 2
- [47] H. Song, S. Yang, and W. Woo. Fast texture transfer for xr avatars via barycentric uv conversion, 2025. 4, 5, 6, 8
- [48] H. Song, B. Yoon, W. Cho, and W. Woo. Rc-smpl: Real-time cumulative smpl-based avatar body generation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 89–98. IEEE, 2023. 1, 2, 4, 5, 6, 8
- [49] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024. 1
- [50] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1
- [51] P. Tran, E. Zakharov, L.-N. Ho, L. Hu, A. Karmanov, A. Agarwal, M. Goldwhite, A. B. Venegas, A. T. Tran, and H. Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *arXiv preprint arXiv:2405.16204*, 2024. 1, 2, 9
- [52] P. Tran, E. Zakharov, L.-N. Ho, A. T. Tran, L. Hu, and H. Li. Voodoo 3d: volumetric portrait disentanglement for one-shot 3d head reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10336–10348, 2024. 1, 2
- [53] X. Tu, L. Radl, M. Steiner, M. Steinberger, B. Kerbl, and F. de la Torre. Vrsplat: Fast and robust gaussian splatting for virtual reality. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(1):1–22, 2025. 1
- [54] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics*, 24(4):1643–1652, 2018. 5
- [55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [56] F. Weidner, J. Orlosky, S. Yoshimoto, K. Kiyokawa, M. Speicher, D. Saredakis, and B. Fröhlich. A systematic review on the visualization of avatars and agents in AR & VR displayed using head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2596–2606, 2023. doi: 10.1109/TVCG.2021.3134637 2
- [57] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011. 6
- [58] E. Wolf, M. L. Fiedler, N. Döllinger, C. Wienrich, and M. E. Latoschik. Exploring presence, avatar embodiment, and body perception with a holographic augmented reality mirror. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 350–359. IEEE, 2022. 5
- [59] W. Yin, Z. Cai, R. Wang, A. Zeng, C. Wei, Q. Sun, H. Mei, Y. Wang, H. E. Pang, M. Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025. 1
- [60] B. Yoon, H.-i. Kim, G. A. Lee, M. Billingham, and W. Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 547–556. IEEE, 2019. 2
- [61] Z. Yu, W. Cheng, X. Liu, W. Wu, and K.-Y. Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16943–16953, 2023. 2
- [62] J. Zhang, Z. Wu, Z. Liang, Y. Gong, D. Hu, Y. Yao, X. Cao, and H. Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5535–5545, 2025. 4, 5, 6, 8
- [63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 6
- [64] Z. Zhao, Z. Bao, Q. Li, G. Qiu, and K. Liu. Psavatar: A point-based morphable shape model for real-time head avatar creation with 3d gaussian splatting. *arXiv preprint arXiv:2401.12900*, 2024. 2
- [65] X. Zheng, C. Wen, Z. Li, W. Zhang, Z. Su, X. Chang, Y. Zhao, Z. Lv, X. Zhang, Y. Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 1, 2
- [66] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4):1–19, 2023. 2
- [67] 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

VRGaussianAvatar: Integrating 3D Gaussian Splatting Avatars into VR

Appendix.

A IMPLEMENTATION DETAILS

A.1 VR System Configuration

The VR runtime consists of a Unity-based VR Frontend and a Python-based GA Backend running on the same local machine. The HMD is connected via a wired link to ensure stable high-bandwidth transmission. Key system parameters are summarized in Table 3.

Table 3: VR System and Rendering Hyperparameters.

Parameter	Value
<i>Rendering Backend (Python)</i>	
Render Resolution (per eye)	1024 × 1024
Field of View (FOV)	65.0°
Inter-pupillary Distance (IPD)	0.055 m
JPEG Compression Quality	80
Binocular Batching	Enabled (Single Pass)
<i>VR Frontend (Unity)</i>	
Target HMD	Meta Quest 3
Connection Type	Wired Link (USB 3.0)
Network Protocol	WebSocket (Localhost)
IK Solver	Final IK (VRIK)
Update Rate	Client-driven (Async)

A.2 Offline Reconstruction Pipeline in GA Backend

We utilize the Large Human Model (LHM-500M) [42] architecture for the single-image 3D Gaussian Splatting reconstruction. This architecture leverages powerful pre-trained encoders, specifically DINOv2 and Sapiens, to extract semantic and geometric features from a single RGB image, which are then decoded into 3D Gaussian parameters via a transformer-based network. For a more detailed description of the reconstruction pipeline and training strategy, please refer to LHM [42]. The detailed hyperparameters for the model architecture and training process are listed in Table 4.

Table 4: Hyperparameters for the Offline Reconstruction Pipeline.

Category	Parameter	Value
Model Architecture	Image Encoder	DINOv2-ViT-L/14 (Frozen)
	Fine Encoder	Sapiens-1B (Frozen)
	Transformer Config	5 Layers, 16 Heads, 1024 Hidden Dim
	Point Embeddings	1024 Dimensions
	Gaussian MLP	2 Layers, 512 Neurons, SiLU Activation
Training	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
	Learning Rate	4×10^{-5} (Cosine Schedule)
	Weight Decay	0.05
	Batch Size	4 (per GPU)
	Total Epochs	60
Loss Weights	Masked Pixel Loss	1.0
	Perceptual Loss	1.0
	Face ID Loss	0.05
	Offset Regularization	1000.0

A.3 Streaming Protocol and Latency

While our reference setup uses localhost, deploying the GA Backend over a remote network can introduce additional latency and bandwidth constraints. Maintaining interactive frame delivery in such settings may therefore require bitrate–quality trade-offs (e.g., stronger compression or lower resolution), depending on network conditions.

We implement a WebSocket-based streaming protocol in the reference setup. The VR Frontend sends per-frame HMD pose and SMPL-X parameters as JSON, and the GA Backend returns rendered stereo images as concatenated JPEG binaries. Table 5 reports the end-to-end latency breakdown measured under this localhost configuration.

Note that these stages run asynchronously in a pipelined manner. As a result, the achievable frame rate is primarily bounded by the GA Backend rendering time (and encoding), while the end-to-end latency reflects the full pipeline delay from tracking to display. We report FPS in Table 2 using the GA Backend throughput (rendering + encoding).

B MORE QUALITATIVE RESULTS AND CHALLENGING CASES

This section provides additional qualitative comparisons of VRGaussianAvatar against baseline mesh avatar representations under identical HMD-only control. We highlight representative regions where our 3DGS avatar preserves fine appearance cues (e.g., facial details and clothing texture), and we include a single challenging case focused on hands. Representative examples are shown in Figure 8.

Hands are particularly difficult to render sharply in our current pipeline. Compared to other body regions, hands have a higher kinematic freedom and undergo larger pose changes during natural interaction. However, the underlying 3DGS hand representation is reconstructed with a finite Gaussian density, which can be insufficient to stably represent fast, highly articulated deformations. As a result, our method can exhibit localized blurring around fingers during motion, and this artifact can be more noticeable than in relatively rigid regions or in mesh-based reconstructions.

As future work, we plan to incorporate a dynamic hand pose representation and integrate a VR-oriented 3DGS hand model that is designed for frequent, highly articulated hand motions.

Table 5: Latency breakdown of the streaming pipeline measured under the localhost configuration. Stages are grouped by where they execute (GA Backend vs. VR Frontend).

Stage	Location	Time (ms)
Tracking & inverse kinematics	VR Frontend	~ 1
Localhost communication (request/response)	GA Backend \leftrightarrow VR Frontend	< 1
Stereo rendering (Binocular Batching)	GA Backend	$\sim 23\text{-}27$
Image encoding (JPEG, multi-threaded)	GA Backend	$\sim 1\text{-}2$
Image decoding & display	VR Frontend	$\sim 5\text{-}10$
End-to-end latency	–	$\sim 30\text{-}41$

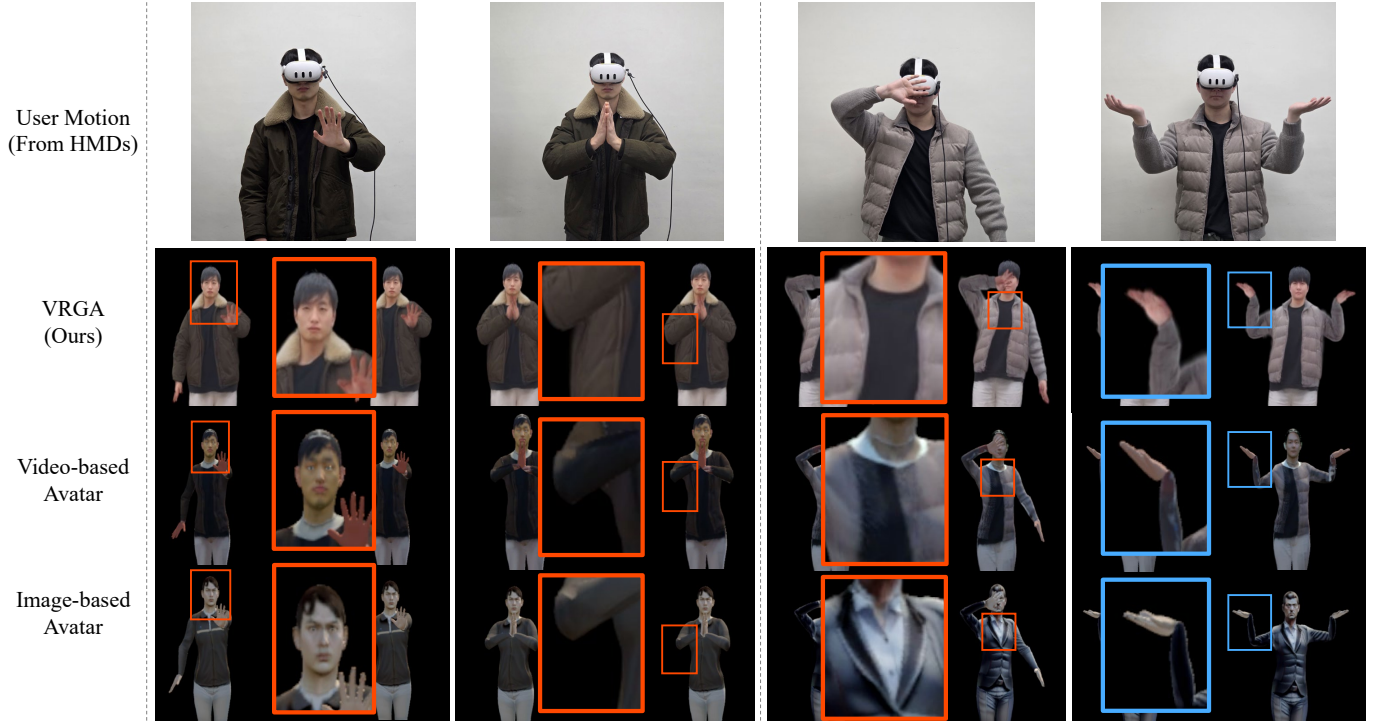


Fig. 8: **More qualitative results.** Orange boxes indicate representative close-ups where our method preserves facial and clothing details. Sky-blue boxes highlight a challenging hand region, discussed in [Appendix B](#).