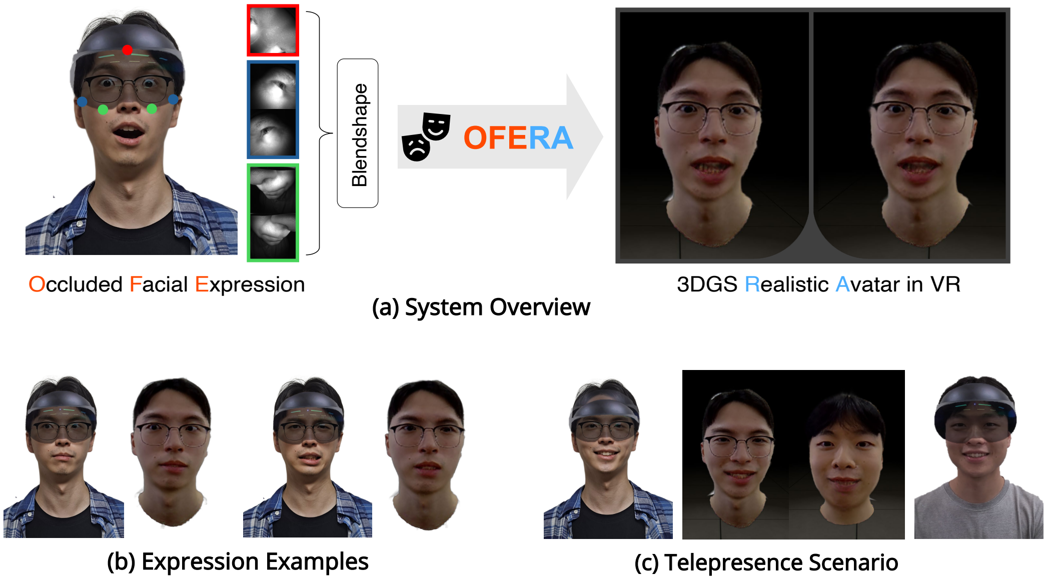
OFERA: Blendshape-driven 3D Gaussian Control for Occluded Facial Expression to Realistic Avatars in VR

Seokhwan Yang, Boram Yoon, Seoyoung Kang, Hail Song, Woontack Woo



(a) System Overview

(b) Expression Examples

(c) Telepresence Scenario

Fig. 1: We introduce **OFERA**, a system that transfers the facial expressions of VR headset users to photorealistic 3D Gaussian Splatting (3DGS) [27] avatars using blendshape data. (a) Blendshape coefficients obtained from a VR headset are mapped by OFERA to animate a 3DGS avatar. (b) OFERA faithfully reproduces diverse facial expressions of VR headset users. (c) The system enables multi-user telepresence in VR, where users interact through realistic 3DGS avatars that convey their expressions. *Note: The IR images shown in (a) are included for visualization only and sampled from the Ava-256 [42] dataset, since raw images from built-in headset cameras are not accessible. Images of users wearing headsets are created by overlaying semi-transparent headset renders on the original face images to illustrate headset-induced facial occlusion.*

Abstract—We propose **OFERA**, a novel framework for real-time expression control of photorealistic Gaussian head avatars for VR headset users. Existing approaches attempt to recover occluded facial expressions using additional sensors or internal cameras, but sensor-based methods increase device weight and discomfort, while camera-based methods raise privacy concerns and suffer from limited access to raw data. To overcome these limitations, we leverage the blendshape signals provided by commercial VR headsets as expression inputs. Our framework consists of three key components: (1) Blendshape Distribution Alignment (BDA), which applies linear regression to align the headset-provided blendshape distribution to a canonical input space; (2) an Expression Parameter Mapper (EPM) that maps the aligned blendshape signals into an expression parameter space for controlling Gaussian head avatars; and (3) a Mapper-integrated Avatar (MiA) that incorporates EPM into the avatar learning process to ensure distributional consistency. Furthermore, OFERA establishes an end-to-end pipeline that senses and maps expressions, updates Gaussian avatars, and renders them in real-time within VR environments. We show that EPM outperforms existing mapping methods on quantitative metrics, and we demonstrate through a user study that the full OFERA framework enhances expression fidelity while preserving avatar realism. By enabling real-time and photorealistic avatar expression control, OFERA significantly improves telepresence in VR communication. A project page is available at https://ysshwan147.github.io/projects/ofera/.

Index Terms—Virtual reality, Gaussian avatar facial expression, Telepresence

1 Introduction

In virtual reality (VR) communication, avatars that faithfully reproduce users' appearance and facial expressions are essential for enhancing the sense of telepresence. Two factors are particularly decisive: (i) how realistically avatars resemble users' facial appearance, and (ii) how accurately and responsively they reflect users' facial expressions in real-time. Prior studies have shown that the realism of avatar appearance and the responsiveness of expressions strongly influence social presence and collaboration outcomes in VR and mixed reality environments [2–4, 10, 31, 49]. These findings highlight that both visual fidelity and expressive capability are crucial for effective avatar-mediated communication.

Early approaches to avatar reconstruction relied on mesh-based

- *Seokhwan Yang, Seoyoung Kang and Hail Song are with KAIST UVR Lab. E-mail: {ysshwan147 | sy1009kang | hail96}@kaist.ac.kr.*
- *Boram Yoon is with KAIST KI-ITC ARRC. E-mail: boram.yoon1206@kaist.ac.kr.*
- *Woontack Woo is with KAIST UVR Lab and KAIST KI-ITC ARRC. Corresponding Author. 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.xxxx/TVCG.201x.xxxxxxx

parametric models that represent identity and expression with low-dimensional parameters [5, 6, 35]. These models enabled efficient animation and controllable expression synthesis, and later works improved them with disentangled representations of pose and expression as well as finer geometry and appearance recovery [12, 15, 68]. While lightweight and interpretable, mesh-based avatars remain visually limited compared to modern rendering techniques.

Neural rendering has since brought a major leap in realism. Neural Radiance Fields (NeRF) [43] first enabled photorealistic novel-view synthesis, and subsequent studies extended this framework to dynamic and controllable head avatars conditioned on expressions or audio signals [18, 21, 22]. Despite their realism, NeRF-based avatars suffer from long training and rendering times, limiting their use in real-time VR communication.

Recently, 3D Gaussian Splatting (3DGS) [27] emerged as an efficient alternative, representing a scene as a collection of Gaussian primitives. This formulation supports interactive frame rates without sacrificing photorealistic detail. Building on this, recent work has demonstrated animatable 3DGS head avatars with improved expression control, relightability, and cross-subject generalization [46, 57, 67]. These advances make 3DGS particularly well-suited for VR communication. Nevertheless, most pipelines still assume unobstructed facial images, which is infeasible when VR headsets occlude large parts of users' faces.

To address this issue, prior works have explored two directions: *non-vision-based* and *vision-based* methods. Non-vision-based approaches [33, 36, 60, 63] employ additional sensors such as EMG, EOG, or microphones, but these increase device weight, reduce comfort, and often suffer from limited accuracy. Vision-based methods [44,45,54,58] instead use cameras to capture partial facial regions, yet they raise privacy concerns due to unintended body capture and face additional technical challenges, such as heterogeneous camera configurations across VR headsets and restricted access to raw data.

We tackle these limitations by exploiting the FACS [14]-based blendshape signals provided by commercial VR headsets. These signals are derived from internal vision pipelines without exposing raw facial images, thereby reducing privacy concerns while still encoding core facial actions. Importantly, blendshapes require no additional sensors and are supported across most consumer VR headsets, ensuring broad applicability. Based on this observation, we present **OFERA** (Occluded Facial Expression to Realistic Avatar), a blendshape-driven framework that enables real-time expression control of photorealistic Gaussian head avatars for VR headset users. An overview of our system is shown in Fig. 1.

Our contributions are threefold:

1. **Expression Parameter Mapper (EPM):** an MLP-based model that maps headset-provided blendshape signals into an expression parameter space for controlling photorealistic Gaussian head avatars, enabling expressive control without raw camera access and addressing privacy and data access constraints.
2. **Mapper-aware Data Adaptation:** a pipeline that aligns headset-specific blendshapes with the EPM distribution through Blendshape Distribution Alignment (BDA) and reduces train–test mismatch by integrating the EPM into avatar training via Mapper-integrated Avatar (MiA).
3. **End-to-End Real-Time Gaussian Avatar Rendering Pipeline:** a complete pipeline that processes VR headset blendshape sensing, expression mapping, and Gaussian avatar updates to achieve real-time stereo rendering of photorealistic avatars in VR environments.

We quantitatively validate that our MLP-based EPM outperforms prior mapping strategies in converting headset-provided blendshape signals into a controllable facial expression parameter space, achieving lower errors at both the parameter level and in vertex-based reconstructions. Beyond numerical evaluation, we conduct a user study in which participants assessed the expressiveness of their own avatars while wearing a VR headset. The results confirm that OFERA enables avatars to faithfully reproduce occluded facial expressions, and further demonstrate the effectiveness of our mapper-aware adaptation pipeline, where BDA reduces distribution mismatch and MiA ensures consistency during avatar training. Taken together, these findings show that OFERA not only advances the accuracy and robustness of expression mapping but also enhances user-perceived realism and expressiveness. By enabling photorealistic Gaussian avatars to reflect VR headset users' expressions in real-time, OFERA facilitates richer emotional exchange and more effective communication in immersive VR environments.

2 Related Work

2.1 Avatars in VR Communication

Avatars play a central role in enabling telepresence in VR environments, where the sense of being together depends on how faithfully avatars convey users' appearance, behavior, and expressions. Early research highlighted how avatars influence social dynamics and communication: Bailenson and Yee [2] introduced the concept of Transformed Social Interaction, demonstrating that avatar-mediated behaviors can strongly affect social presence, while Slater et al. [49] showed that embodying a virtual body can induce a strong sense of self-representation and agency.

Subsequent studies examined how the realism and design of avatars shape user experience in communication. Bailenson et al. [3] investigated the effect of behavioral and form realism of real-time avatar faces, showing that more realistic facial rendering enhances perceived social presence. Baker et al. [4] analyzed avatar-mediated communication in social VR and emphasized the importance of responsive avatars for interaction quality. More recent work [10] explored the role of avatar heads and facial expressions in mixed reality collaboration, demonstrating that they significantly influence perceived social presence. Similarly, Lee et al. [31] reported that avatar visibility affects joint agency in collaborative tasks.

Other studies further investigated how individual and contextual factors modulate avatar-mediated communication. Research has shown that gender differences influence the perception of avatar faces and interpersonal distance [24], while collaboration context and personality traits also shape user experience in social VR [26]. In augmented and mixed reality settings, avatar appearance, body part representations, and visual fidelity have been shown to affect social presence, including avatar appearance [64], hand models [65], avatar transparency [66], and emotion-based prioritized facial expressions [25]. Beyond head-only representations, full-body generation approaches such as RC-SMPL [52] demonstrated in user studies that avatar realism influences embodiment and user experience in VR.

Together, these studies highlight that avatar design critically impacts social presence and telepresence in VR and mixed reality communication. Building on these insights, we focus on enhancing avatar expressiveness by enabling realistic facial expression control for VR headset users through our proposed framework, OFERA.

2.2 Head Avatar Reconstruction

Realistic reconstruction of head avatars has long been a central research topic in computer vision and graphics, particularly for applications in virtual communication. The key goal is to create avatars that resemble users' identities while remaining animatable to capture facial expressions, which are essential for natural interaction.

The 3D Morphable Model (3DMM) by Blanz and Vetter [5] introduced a statistical framework for facial shape and texture modeling using PCA, enabling reconstruction from monocular images. FaceWarehouse [6] further expanded this direction with a large-scale dataset supporting identity–expression bilinear modeling. Li et al. later proposed FLAME [35], which disentangles identity, pose, and expression in a unified model with articulated jaw, neck, and eyes, and MICA [69] improved the encoding of identity with high-quality reconstruction. Other studies extended these models with unsupervised regression, hybrid losses, and affective expression modeling [11, 12, 15, 51, 68]. Mesh-based models thus provide compact and expressive control for avatar reconstruction and animation, and their low-dimensional parameters are often used to drive more detailed neural representations.

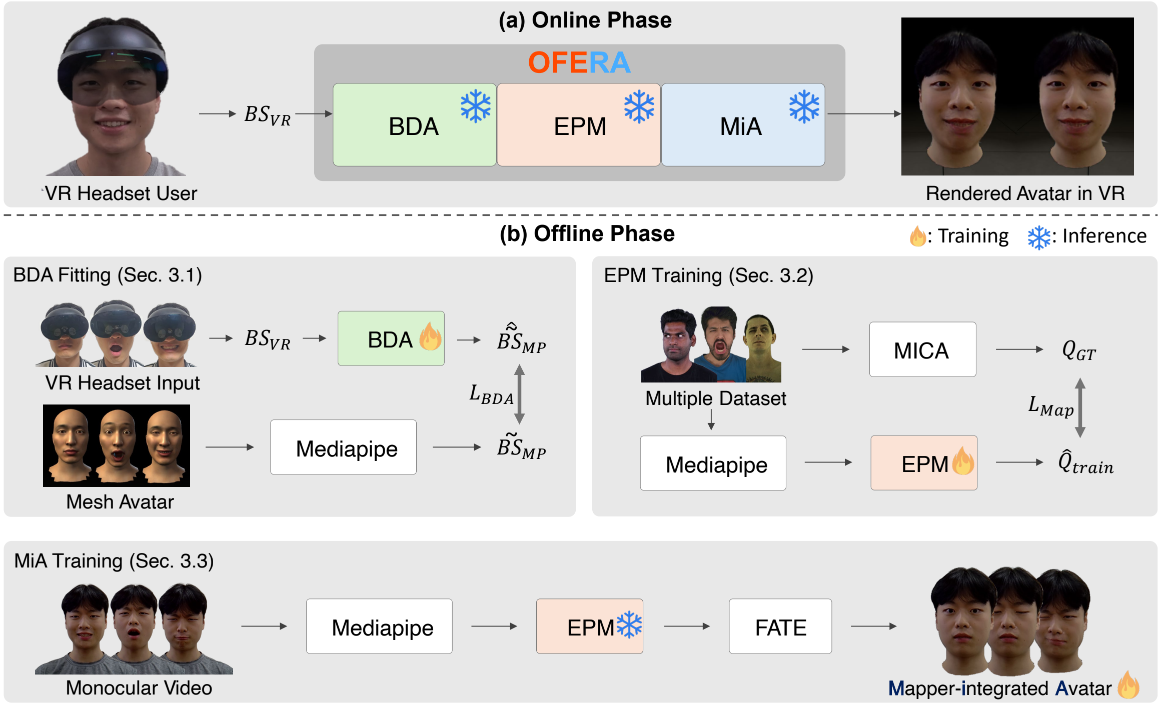


Fig. 2: Overview of OFERA, which operates in online and offline phases. (a) In the online phase, headset blendshapes are processed through BDA, EPM, and MiA to drive a photorealistic 3D Gaussian avatar in real-time. (b) The offline phase involves fitting BDA (Sec. 3.1) and training EPM (Sec. 3.2) and MiA (Sec. 3.3) to ensure accurate expression mapping and consistent avatar rendering across headsets.

While mesh-based models provide efficient control, they are inherently limited in realism and fine-scale detail. Neural rendering has recently enabled photorealistic and animatable avatar reconstruction. NeRF [43] pioneered neural radiance fields for novel view synthesis, and subsequent works such as NerFace [18], AD-NeRF [22], and Neural Head Avatars (NHA) [21] extended this framework to dynamic and controllable head avatars. Additional studies further incorporated audio conditioning, parametric priors, or efficiency improvements [13, 53, 61, 70]. These methods achieved highly realistic avatars but remain impractical for real-time VR communication due to slow training and rendering. To address these limitations, 3D Gaussian Splatting (3DGS) [27] emerged as an efficient alternative, supporting real-time rendering with high fidelity. Gaussian Head [57] demonstrated its applicability for head avatars, and GaussianAvatars [46] generalized the approach across subjects. Many other extensions have advanced expression control, relightability, and efficiency [9, 19, 23, 32, 34, 40, 48, 50, 56, 67], showing that 3DGS can represent dynamic heads with fine detail while supporting interactive frame rates, making them well-suited for VR communication scenarios.

In our work, we adopt FATE [67] as the backbone model. FATE achieves state-of-the-art reconstruction quality and includes a full-head completion module that synthesizes plausible side and back views from monocular input, an essential property for VR avatars. We further integrate our Expression Parameter Mapper into the training process to enable consistent expression control.

2.3 Facial Expression Capture with VR Headsets

When users wear a VR headset, much of the face is occluded, making expression capture challenging. Prior work has explored both non-vision-based and vision-based methods.

Non-vision-based methods employ additional sensors to capture signals correlated with muscle activity or skin deformation. For example, AUGlasses [36] and IMUFace [63] use inertial sensing, BioFace-3D [60] leverages EMG/EOG, and EyeEcho [33] applies ultrasonic sensing. Others explored RF antennas or body-mounted cameras [7,8,28]. These approaches avoid direct imaging but often increase weight, require skin contact, and lack robustness in noisy environments. Vision-based methods integrate cameras into headsets to capture visible regions. Olszewski et al. [44] regressed avatar parameters from headset cameras, FaceVR [54] combined RGB-D and IR sensing for reenactment, and Wei et al. [58] and Patel et al. [45] advanced photorealistic facial animation. Other works further improved fidelity with codec avatars and appearance models [1, 38]. Beyond direct camera sensing, VOODOO XP [55] demonstrated that headset-provided blendshapes can be retargeted to a generic rigged mesh avatar and then used to drive photorealistic head reenactment for VR telepresence. Despite their robustness, these methods suffer from device-specific constraints and privacy issues since they require raw facial video streams.

In contrast, our system leverages blendshape parameters released by commercial VR headsets. Derived from internal vision pipelines, they preserve essential facial actions while avoiding privacy concerns. In particular, ARKit blendshapes are FACS [14]-based and widely supported, making them an accessible and generalizable input for our framework, OFERA. Prior works have also attempted matrix-based or linear mappings from blendshapes to parametric model expressions [37, 62], but these approaches are limited in expressiveness and generalization, which motivates our design of a learning-based Expression Parameter Mapper.

3 Method

Our framework, **OFERA** (Occluded Facial Expression to Realistic Avatar), provides an end-to-end pipeline that converts headset-derived blendshapes into photorealistic Gaussian head avatars in VR. As shown in Fig. 2, headset blendshapes are first aligned to a canonical distribution by the Blendshape Distribution Alignment (BDA) module, then mapped to a continuous expression parameter space by the Expression Parameter Mapper (EPM), which is instantiated using the FLAME [35] model in this work. The Mapper-integrated Avatar (MiA) ensures that avatars are trained and rendered consistently with the EPM outputs, and the resulting Gaussian updates are streamed to a stereo rendering pipeline in real-time. In this way, OFERA bridges device-dependent blendshape inputs with high-fidelity 3D Gaussian avatars, enabling natural and consistent expressions in VR communication.

3.1 Blendshape Distribution Alignment (BDA)

We use ARKit[1]-based blendshapes as the primary input to our framework. Since they are based on the Facial Action Coding System (FACS) [14], ARKit blendshapes can represent key facial expressions and are widely supported across various VR headsets. Importantly, these blendshape signals are derived from internal headset pipelines and can be accessed without exposing raw facial images, making them well-suited for privacy-preserving facial expression capture under headset-based sensing. However, ARKit blendshapes were originally designed for use with webcams or RGBD cameras, and the way different headsets provide ARKit-compatible coefficients is not standardized. For example, the Meta Quest Pro generates OVR[2] blendshapes from built-in sensors and then simply maps them to ARKit parameters with predefined scalar weights (e.g., 0.5, 0.75, or 1.0). Such coarse remapping is sufficient for basic animation, but introduces significant discrepancies when these transformed ARKit values are used in learning-based pipelines such as our EPM, leading to increased prediction errors. As a result, the raw values obtained from VR headsets (BS_{VR}) are not directly compatible with the EPM trained on Mediapipe [39]-derived values.

To resolve this discrepancy, we introduce **Blendshape Distribution Alignment (BDA)**, a process that transforms VR headset blendshapes into a distribution suitable for EPM inference. Training BDA requires paired data $(BS_{VR}, \tilde{BS}_{MP})$, where BS_{VR} are VR headset–driven ARKit blendshapes and $\tilde{BS}_{MP}$ are Mediapipe-style ARKit blendshapes.

Constructing such pairs directly is challenging, since a participant wearing a VR headset cannot provide unobstructed facial images for Mediapipe analysis. To overcome this, we built a pseudo paired dataset, which serves as a proxy for real paired samples; details are provided in the supplementary material. For each subject, we generated a mesh avatar using a commercial framework[3], a tool that creates a personalized 3D mesh model from a few facial photographs. The generated avatar has a facial geometry and appearance closely resembling the subject (e.g., face shape and facial features) and comes with a pre-rigged set of blendshapes for expression control. While the subject performed various expressions with the headset on, the corresponding blendshapes (BS_{VR}) were recorded via the Meta Movement SDK[4]. These blendshapes were then used to animate the mesh avatar, and Mediapipe was applied to the rendered frontal images to extract Mediapipe-style ARKit blendshapes, which we denote as $\tilde{BS}_{MP}$. This indirect strategy enabled us to approximate the required $(BS_{VR}, \tilde{BS}_{MP})$ pairs offline. The overall BDA fitting process is illustrated in Fig. 2.

We note that the blendshape rig and amplitude calibration of the commercial mesh avatar are not guaranteed to be strictly consistent with those used by the VR headset. As a result, the constructed pseudo paired data may contain residual inter-model bias. Accordingly, BDA is not intended to establish a perfect semantic correspondence between headset-provided and Mediapipe-derived blendshapes, but rather to serve as a practical calibration step that mitigates large-scale distribution mismatch across heterogeneous input sources.

Using the constructed dataset, we fit a linear regression model to align BS_{VR} with $\tilde{BS}_{MP}$:

$$\hat{\tilde{BS}}_{MP} = WBS_{VR} + b \tag{1}$$

where $W \in \mathbb{R}^{51\times 51}$ and $b \in \mathbb{R}^{51}$ are regression parameters optimized by minimizing

$$L_{BDA} = \frac{1}{N}\sum_{i=1}^{N} \|\tilde{BS}^{(i)}_{MP} - \hat{\tilde{BS}}^{(i)}_{MP}\|^2 \tag{2}$$

Since the regression model is lightweight and linear, BDA can be executed as a pre-processing module during inference with negligible overhead (< 1 ms per frame). This ensures consistent expression recognition in VR environments and allows the same EPM to be reused across different input sources.

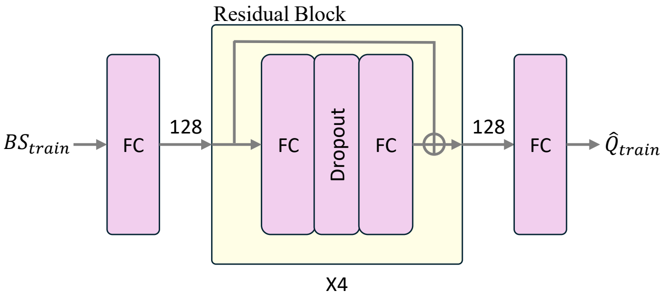


Fig. 3: Architecture of the proposed EPM model. The network takes blendshape parameters (BS_{train}) as input and produces predicted target parameters ($\hat{Q}_{train}$) as output during training. It consists of an initial fully connected (FC) layer, followed by four residual blocks (each with two FC layers and an intermediate dropout layer), and a final FC layer. All FC layers are followed by batch normalization and ReLU activation, except for the last output FC layer. Each hidden FC layer has 128 units.

3.2 Expression Parameter Mapper (EPM)

To represent the facial expressions of VR headset users in a photorealistic Gaussian head avatar in real-time, we design an **Expression Parameter Mapper (EPM)** that maps headset-provided blendshape signals obtained in Sec. 3.1 into a continuous expression parameter space suitable for controlling deformable Gaussian head avatars. In this work, we instantiate this target expression parameter space using the FLAME [35] model, and predict its expression parameters along with jaw and eye pose parameters, which are widely used in facial expression representation and compatible with existing avatar deformation pipelines.

Blendshapes represent human facial expressions as a combination of multiple parameters bounded within $[0, 1]$. Different parameter combinations can yield visually similar expressions, which introduces ambiguity. Moreover, many parameters are strongly correlated (e.g., smiling involves simultaneous motion of lips, eyes, and cheeks), leading to complex dependencies. These properties make blendshape data highly non-linear. This implies that the transformation from ARKit blendshape coefficients to FLAME parameters cannot be sufficiently modeled as a simple linear regression problem. To address this, we adopt a mapping strategy based on a Multi-Layer Perceptron (MLP) architecture, as illustrated in Fig. 3. This design enables stable and accurate mapping of user expressions into the FLAME parameter space with low computational cost, allowing the system to operate without noticeable latency in VR headset environments.

For training the EPM, we use frontal face video datasets, as illustrated in Fig. 2. For each frame, ARKit blendshape values $BS_{train} \in [0, 1]^{51}$ are extracted using Mediapipe [39], and ground-truth FLAME parameters $Q_{GT} \in \mathbb{R}^{68}$ are obtained using the MICA [69] model. The ground-truth parameters Q_{GT} consist of FLAME expression parameters $e \in \mathbb{R}^{50}$ and the 6D rotations of the jaw, left eye, and right eye $(r_{jaw}, r^{L}_{eye}, r^{R}_{eye}) \in \mathbb{R}^{18}$.

As shown in Eq. (3), the EPM transforms BS_{train} into predicted parameters $\hat{Q}_{train}$, and training is performed by minimizing the L1 loss L_{Map} between $\hat{Q}_{train}$ and Q_{GT} as defined in Eq. (4):

$$\hat{Q}_{train} = EPM(BS_{train}) \tag{3}$$

$$L_{Map} = \|\hat{Q}_{train} - Q_{GT}\|_1 \tag{4}$$

[1] Apple Developer, "ARFaceAnchor.BlendShapeLocation", https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation

[2] Meta Horizon, "Blendshape Visual Reference for Movement Extensions for OpenXR", https://developers.meta.com/horizon/documentation/native/android/move-ref-blendshapes

[3] Avaturn, "Avaturn", https://avaturn.me

[4] Meta Horizon, "Movement SDK for Unity - Overview", https://developer.oculus.com/documentation/unity/move-overview

Since the model must generalize across diverse identities and a wide range of expression variations, we construct the training dataset by merging multiple sources and apply subject-wise sampling to mitigate distribution bias. This training process enables the EPM to effectively capture a broad distribution of facial expressions and generalize to unseen users.

3.3 Mapper-integrated Avatar (MiA)

Existing FLAME [35]-based avatar training frameworks typically extract FLAME parameters from an input video, render the Gaussian avatar accordingly, and optimize the model by comparing the rendering with the ground-truth image. This approach inherently depends on the distribution of the extracted FLAME parameters, and is therefore strongly influenced by the parameter estimation method (e.g., MICA [69]).

In our system, we adopt the FATE [67] model as the backbone for Gaussian avatar generation. FATE is particularly suitable for our setting, as it not only achieves high-quality novel view rendering but also incorporates a *full-head completion* module that synthesizes plausible side and back views, which is crucial for VR scenarios. However, since FATE is optimized on the parameter distribution produced by MICA, any mismatch in parameter distribution during inference inevitably degrades rendering quality.

In practice, our framework relies on ARKit blendshape coefficients obtained from a frontal user video, denoted as BS_{vid}, which are converted into FLAME parameters by the EPM. The output distribution of the EPM, however, shows a discrepancy with that of MICA. Directly applying an avatar trained on MICA-based parameters to EPM outputs would amplify this mismatch, resulting in degraded expression rendering.

To address this issue, we propose the **Mapper-integrated Avatar (MiA)** strategy, where the EPM is directly integrated into the avatar training pipeline, as illustrated in Fig. 2. Specifically, for each frame used during avatar training, ARKit blendshapes BS_{vid} are passed through the EPM to obtain predicted FLAME parameters:

$$\hat{Q}_{vid} = EPM(BS_{vid}) = (\hat{e}, \hat{r}_{jaw}, \hat{r}^{L}_{eye}, \hat{r}^{R}_{eye}) \tag{5}$$

These parameters are then used to deform the FLAME mesh following the formulation:

$$T(\hat{Q}_{vid}) = LBS\Big(B_P(\Theta; \{\hat{r}_{jaw}, \hat{r}^{L}_{eye}, \hat{r}^{R}_{eye}\} + \Delta P) \; + \; B_E(\Psi; \hat{e} + \Delta E)\Big) \tag{6}$$

where LBS denotes the linear blendshape skinning function, B_P and B_E denote the pose- and expression-dependent blendshapes of FLAME, respectively.

As a result, the loss terms originally defined in FATE are adapted such that the image reconstruction loss L_{L1}, perceptual loss L_{vgg}, and FLAME regularization L_{flame} are all computed with respect to $T(\hat{Q}_{vid})$. We denote these modified terms as L'_{L1}, L'_{vgg}, and L'_{flame}. The complete training objective of MiA is thus given by:

$$L = L'_{L1} + \lambda_1 L'_{vgg} + \lambda_2 L_{lap} + \lambda_3 L'_{flame} + \lambda_4 L_{scale} \tag{7}$$

Here, L_{lap} denotes Laplacian smoothing of mesh vertices, and L_{scale} enforces constraints on Gaussian anisotropy. These regularization terms are independent of $\hat{Q}_{vid}$ and remain identical to those in the original FATE framework.

By aligning the avatar training process directly with the distribution of EPM outputs, the resulting Gaussian avatar can faithfully and consistently reflect users' expressions when driven by VR headset inputs at inference time.

3.4 End-to-End Real-Time Gaussian Avatar System in VR

To enable end-to-end, real-time control of photorealistic Gaussian head avatars under headset-only sensing, we build a unified system that integrates lightweight communication between a Python server and a Unity client, along with expression mapping, avatar updates, and VR rendering. During initialization, once the connection is established, a Gaussian avatar with a canonical expression is transmitted from the server to the client, and the Unity client performs stereo rendering of this idle avatar. In the execution stage, ARKit blendshape data BS_{VR} are extracted from the VR headset via the Meta Movement SDK and sent to the server. Upon receiving the blendshapes, the server processes them through the BDA, EPM, and MiA modules to estimate the optimal FLAME [35] parameters $\hat{Q}_{VR}$. These parameters are then used to update the Gaussian avatar, where each Gaussian [27]'s position, rotation, and scale are modified according to the expression changes. Since opacity and spherical harmonics (SH) coefficients remain constant and are independent of expression changes, they are transmitted only once during initialization, and subsequent updates transmit only position, rotation, and scale. This optimization significantly reduces communication overhead and improves performance, enabling real-time interaction in VR environments.

The Unity client supports dynamic updates of the Gaussian avatar by applying streamed expression-driven changes to Gaussian attributes and renders the avatar with the user's expressions in real-time within the VR environment. The overall end-to-end latency from blendshape acquisition to avatar rendering is approximately $\sim$20ms, which is sufficient for interactive VR applications. This system-level design enables practical deployment of Gaussian avatars in VR by bridging headset-derived expression signals and real-time avatar rendering, rather than introducing new avatar representations.

4 Experiment

4.1 Experimental Setup

Evaluation Design. Evaluating the proposed OFERA system requires both quantitative and qualitative assessments, as well as a user study. Since OFERA is designed to animate Gaussian head avatars in VR environments where users wear headsets, a ground-truth (GT) facial mesh captured under the same conditions is not accessible. Therefore, we distinguish two types of evaluation: (1) for the **EPM**, we perform quantitative and qualitative evaluation using front-facing facial images, from which ARKit blendshapes can be extracted via MediaPipe [39]; (2) for **BDA** and **MiA**, objective evaluation is infeasible, and we rely on a user study to validate their effectiveness in VR environments. Additionally, to evaluate the overall OFERA pipeline, we compare our system against baseline mapping approaches within the user study.

Baselines. To validate the effectiveness of our MLP-based EPM module, we compare it with two existing approaches for mapping ARKit blendshapes to FLAME [35] parameters. Liu et al. [37] introduced a **matrix-based mapping** method, where a fixed 51×103 matrix transforms 51-dimensional ARKit blendshapes into 100 expression parameters and 3 jaw rotations of the FLAME model via direct multiplication. This approach is simple and efficient but limited by its inability to capture non-linear dependencies. Yan et al. [62] proposed a **ridge regression-based linear mapping**, where ARKit blendshapes are linearly projected to FLAME parameters, including 100 expressions, 3 jaw rotations, and 6 eye rotations. While the ridge regularization improves stability, the linear model cannot represent the complex interactions among blendshape parameters. These two mapping strategies serve as baselines for evaluating our EPM.

Dataset. To train the EPM module across diverse distributions of facial identity and expression, we combine multiple datasets: INSTA [70], NeRSemble [29], and Ava-256 [42]. From each dataset, we randomly select 10 subjects, resulting in 30 subjects in total. Each dataset is split into 8 subjects for training, 1 for validation, and 1 for testing, ensuring subject-disjoint partitions and preventing distributional bias across datasets. We specifically use front-facing images of the test subjects for evaluation: ground-truth FLAME [35] parameters are obtained using the MICA [69] tracker, and ARKit blendshapes are extracted from the same frames using MediaPipe [39]. The extracted blendshapes are then mapped into FLAME parameters using each baseline and our EPM, enabling direct comparison against the ground-truth FLAME results. The detailed dataset statistics are provided in the

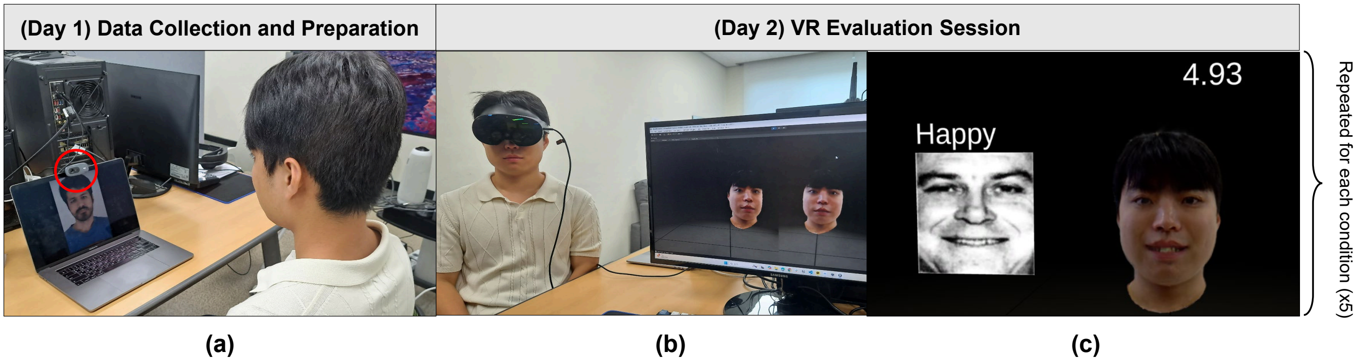

(Day 1) Data Collection and Preparation
(Day 2) VR Evaluation Session
Happy
4.93
Repeated for each condition (x5)
(a)
(b)
(c)

Fig. 4: Overview of the two-day user study for facial data collection and VR-based evaluation (Day 1 and Day 2). (a) Facial expression recording for data collection and preparation; (b) VR setup for the evaluation session; and (c) an example study task in which participants controlled avatar facial expressions in VR using visual references. Avatar control and post-task questionnaires were repeated for each experimental condition (×5).

supplementary material.

Metrics. We evaluate the accuracy of expression mapping by computing the Root Mean Square Error (RMSE) between the predicted and ground-truth FLAME [35] parameters. In addition, since all FLAME meshes share consistent topology and coordinate systems, we further compute the vertex-wise position error as RMSE. For quantitative evaluation, eye pose parameters were excluded from all methods, as the matrix-based mapping baseline does not predict eye-related parameters. Accordingly, errors were computed using only expression and jaw parameters to ensure a fair comparison across methods. To facilitate qualitative analysis, we visualize the vertex error distribution as heatmaps, allowing us to inspect localized differences in facial geometry. This evaluation protocol enables us to assess not only the parameter-level accuracy of expression mapping but also the geometric fidelity of reconstructed meshes.

4.2 User Study

We conducted a subjective user evaluation to assess the effectiveness of the proposed system in reflecting users' actual facial expressions on a virtual avatar. Specifically, the study examined the extent to which avatar facial expressions were realistically conveyed in VR environments and whether facial parts occluded by a headset could also be accurately represented. To this end, five expression mapping methods were set as experimental conditions to compare with our system, including two baselines and two ablations, apart from ours. For each condition, we generated a realistic avatar model based on the participant's real face and applied each mapping method. In all conditions, participants wore a headset and evaluated the facial expressions of their self-representative avatars in a VR environment. All user study experiments were conducted using a Meta Quest Pro headset.

Study Design and Task. The user experiment was designed with expression mapping method (*method*) as the experimental factor, and five conditions were derived: (1) Matrix-based mapping (*matrix*), (2) Linear-based mapping (*linear*), (3) Ours without BDA (*w/o BDA*), (4) Ours without MiA (*w/o MiA*), and (5) OFERA (*Ours*). Among them, (1) *matrix* and (2) *linear* were set as baseline conditions for direct comparison with our system, while (3) *w/o BDA* and (4) *w/o MiA* were configured as ablation model conditions. The *method* factor was treated as a within-subject factor, such that each participant experienced all five conditions. To reduce potential order effects, the presentation order of the five conditions was counterbalanced using a balanced Latin Square method.

For the evaluation, participants were asked to perform a task while wearing a VR headset: They performed facial expressions by imitating given reference materials (images or short video clips). The avatar's facial expressions changed in real-time according to the participant's actual expressions, and participants observed and evaluated their self-representative avatar. Each avatar was generated in advance by scanning the participant's face so that the avatar closely resembled their own appearance. To allow participants to concentrate solely on facial movements and expression, only a head avatar was used. The reference stimuli consisted of emotion-related images and short video clips, each presented for a fixed duration to allow consistent and sufficient observation. The overall task sequence consisted of two parts: (1) trying to imitate universal facial expressions raised by Paul Ekman[5], and (2) trying to imitate more detailed partial expressions. The latter was designed to examine facial actions in specific parts of the face, informed by the Facial Action Coding System (FACS) [14], which was developed to categorize facial muscle actions associated with emotions. Among FACS, we selected emotion-related action units and also included commonly observed expressions—such as eye movements (e.g., eyes, brow, lid), cheek, nose, lips, and jaw movements—to better capture natural facial behaviors. All participants experienced the same sequence in the order of (1) universal expressions followed by (2) partial expressions, and were thus exposed to a wide variety of expressions for evaluation.

Dependent Variables. To evaluate how well a participant's own expressions were reflected in their avatar and how controllable the expressions were while wearing a VR headset, we employed Sense of Embodiment (SoE) as the main dependent variable, selecting two widely used measurements from previous studies investigating avatar embodiment and facial expressions in virtual environments [20, 30, 59]. First, we adopted the Virtual Embodiment Questionnaire (VEQ) proposed by Roth and Latoschik [47]. Among its subscales, we used Virtual Body Ownership (VBO), measuring the extent to which users perceived the virtual body as their own, and Agency (AG), evaluating the sense of control over the virtual body. Second, we employed the extended VEQ (VEQ+) measurement [16] to capture Self-Identification (SI), which reflects identifying the virtual representation as oneself. We assessed two SI-related subscales: Self-Attribution (SA) and Self-Similarity (SS), both indicating how strongly the avatar is perceived to represent the user in terms of personalization and similarity. VBO and AG from VEQ, and SA and SS from VEQ+ comprised a total of 16 items (4 items each).

Next, we used the Virtual Human Plausibility Questionnaire (VHPQ) to assess the naturalness and coherence of avatars in the VR scene [41]. Specifically, we included the Appearance and Behavior Plausibility (ABP) subscale (6 items) to evaluate the plausibility of the avatar's visual appearance and motion behavior. Finally, the Facial Animation Realism (Real) measurement was adopted based on prior studies [17, 24, 25], consisting of 4 items that assess the realism and naturalness of facial expressions and movements. All questionnaire items were rated on a 7-point Likert scale, and subscales with low relevance or items unmeasurable due to the study setup and purpose were excluded.

[5] Paul Ekman Group, "Universal Facial Expressions", https://www.paulekman.com/resources/universal-facial-expressions/

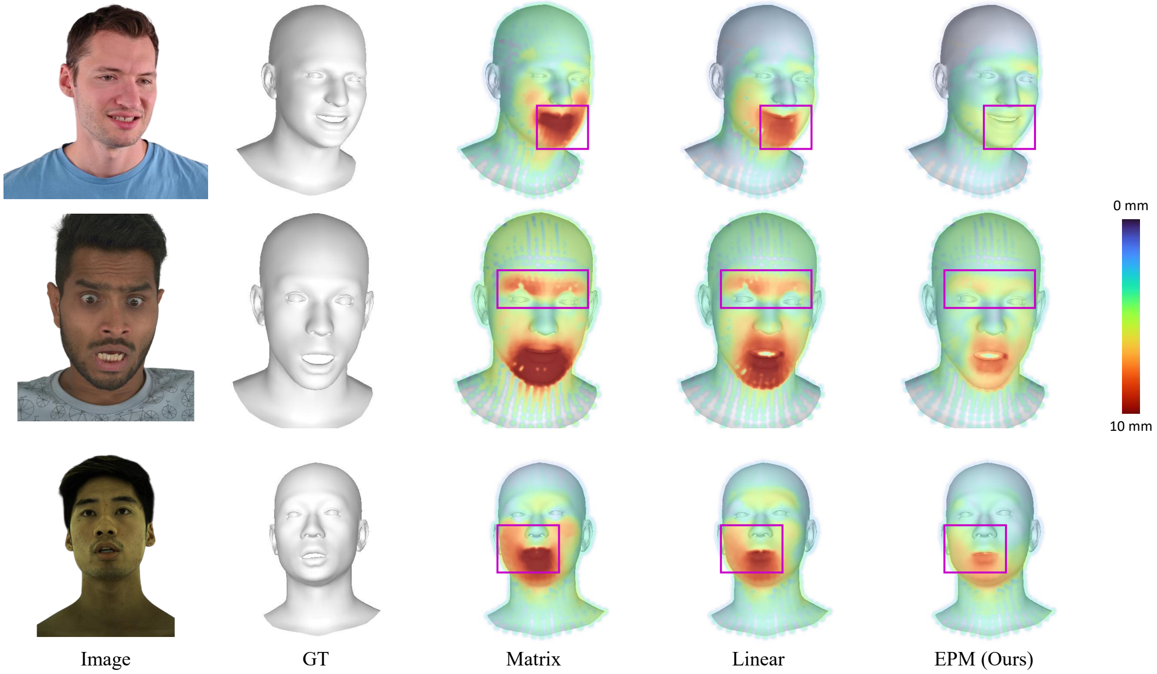


Fig. 5: Qualitative comparison of vertex-wise reconstruction errors visualized as heatmaps. Matrix-based and linear mapping baselines show large errors around expressive regions such as the mouth and eyes, whereas our EPM significantly reduces these localized errors, leading to more faithful reproduction of facial expressions.

At the end of the experiment, we conducted open-ended interviews to gather subjective feedback. The questions focused on which of the five avatar models participants found most similar to their own face and expressions, most expressive, most natural, and realistic. Participants were also asked which avatar they would prefer to use as their own representation and which version they considered the best overall.

Procedure. Prior to the experiment, approval for the study content and procedure was obtained from the Institutional Review Board (IRB). The study consisted of two phases, which were conducted on separate days with at least a one-day interval between them. In the first phase, participants visited for the collection of facial expression data required for avatar reconstruction. During this session, videos of participants performing various facial expressions were recorded. After that, they were given a detailed explanation about the study's purpose and procedure, followed by informed consent and a demographic questionnaire.

In the second phase, participants completed the main task of evaluating avatar facial expressions and movements. At the beginning of this session, the experimenter briefly explained the procedure, and participants were seated and wore a VR headset. Following the on-screen instructions, they tried to make the same facial expressions as those shown in reference images or videos while observing their avatar in VR. An overall illustration of the experiment is shown in Fig. 4.

The five facial expression mapping methods were presented in a counterbalanced order, and thus the same task was repeated for all five conditions. After each method condition, participants completed a post-task questionnaire based on their observations. Once all conditions were completed, they removed the VR headset and participated in a post-experiment interview, where they shared their overall feedback on each method.

Participants. We recruited 20 participants (10 male, 10 female) through the university's community and website. The participants' mean age was 27.8 years (*SD* = 3.17): 27.2 (*SD* = 2.94) for males and 28.4 (*SD* = 2.44) for females. As the study involved observing avatar facial expressions in VR, we also asked about participants' prior experiences with related technologies: Regarding avatar-mediated technologies or applications, 16 reported less than four prior uses (80%), among whom five had no experience at all (25%), while four participants reported more than ten prior uses (20%). For AR/VR-related experiences (e.g., wearable devices, applications), four participants reported less than four prior uses, whereas 16 had more than five experiences, including 11 who had used them more than ten times (55%).

Table 1: Quantitative comparison of expression mapping methods. Our EPM consistently achieves the lowest parameter- and vertex-level errors, demonstrating the benefit of modeling non-linear dependencies between ARKit blendshapes and FLAME parameters.

Mapping Method	**Param Error ↓**	**Vertex Error (mm) ↓**
Matrix	1.038	2.826
Linear	0.573	1.848
EPM (Ours)	**0.505**	**1.593**

5 Results

5.1 Expression Mapping Evaluation

Table 1 summarizes the quantitative performance of different expression mapping methods. Among all methods, the matrix-based mapping approach yields the highest errors. The ridge regression-based linear mapping improves performance over the matrix-based approach. In contrast, our proposed EPM achieves the lowest errors in both parameter space (0.505) and vertex space (1.593), outperforming both baselines. Beyond these quantitative metrics, Fig. 5 visualizes the vertex-wise error distributions across different methods. Consistent with the quantitative results, the matrix-based and linear mappings exhibit substantial deviations around highly expressive regions such as the mouth and eyes, whereas our EPM shows reduced errors in these regions. This trend is particularly noticeable in expressions involving large facial deformations.

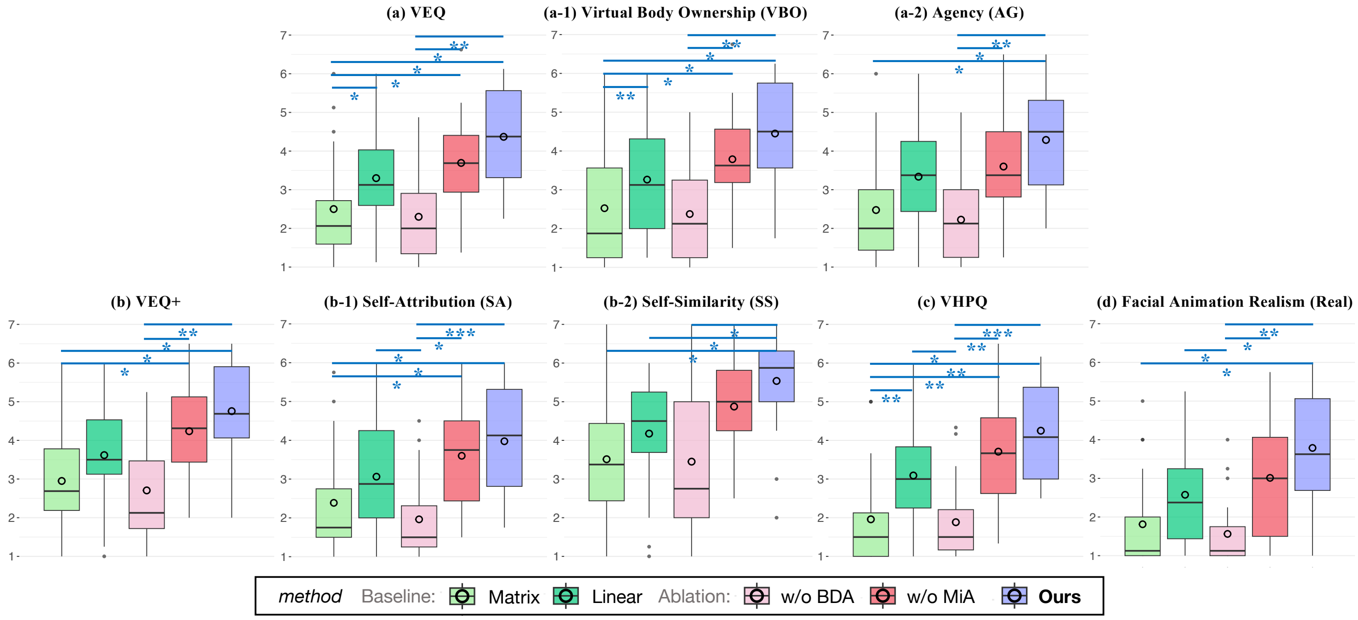


Fig. 6: Results for (a)–(a-2) Virtual Embodiment Questionnaire and subscales; (b)–(b-2) VEQ+ and the subscales associated with Self-Identification; (c) Virtual Human Plausibility Questionnaire; (d) Facial Animation Realism.

5.2 User Study Analysis

Because the user study involved subjective measures across five within-subject conditions, we used a Friedman test for non-parametric analysis ($\alpha = .05$). Post-hoc pairwise comparisons were conducted using a Wilcoxon Signed Rank test with Bonferroni correction. Effect sizes are reported using Kendall's W, with values of approximately .10, .30, and .50 indicating small, moderate, and large effects, respectively. The reliability of the Likert-scale items was verified using Cronbach's alpha. The statistical results are illustrated in Fig. 6.

VEQ: The internal consistency of VEQ and the two subscales (VBO and AG) all showed an acceptable range ($\alpha_{VEQ} = .951$; $\alpha_{VBO} = .901$; $\alpha_{AG} = .947$). A significant main effect of *method* on VEQ scores was found ($\chi^2(4) = 33.841$, $p < .001$, $W = .423$). Pairwise comparisons revealed significant differences between *Ours* and *matrix* ($p = .011$), and *Ours* and *w/o BDA* ($p = .002$). Additional differences revealed in the pairs of *matrix* and *linear* ($p = .046$), *matrix* and *w/o MiA* ($p = .025$), and *w/o BDA* and *w/o MiA* ($p = .022$). For VBO, the main effect of *method* was also significant ($\chi^2(4) = 27.969$, $p < .001$, $W = .350$). Post-hoc revealed significant contrasts in the pairs of *Ours* and *matrix* ($p = .018$), and *Ours* and *w/o BDA* ($p = .003$). Additional pairwise differences were observed for *matrix* and *linear* ($p = .009$), *matrix* and *w/o MiA* ($p = .033$), and *w/o BDA* and *w/o MiA* ($p = .023$) pairs. Regarding the subscale AG, a significant effect of method was also observed ($\chi^2(4) = 30.175$, $p < .001$, $W = .377$). During the post-hoc analysis, significant differences were found between *Ours* and *matrix* ($p = .022$), and *Ours* and *w/o BDA* ($p = .003$), along with a difference for *w/o BDA* and *w/o MiA* ($p = .031$). All other pairs did not significantly differ on VEQ, VBO and AG scores.

VEQ+: The VEQ+ scale and its subscales demonstrated high reliability ($\alpha_{VEQ+} = .959$; $\alpha_{SA} = .932$; $\alpha_{SS} = .975$). The Friedman test indicated a significant effect for method on VEQ+ ($\chi^2(4) = 26.588$, $p < .001$, $W = .332$). In post-hoc analysis, significant differences were found between *Ours* and *matrix* ($p = .010$), and *Ours* and *w/o BDA* ($p = .003$). Further contrasts appeared for *matrix* and *w/o MiA* ($p = .028$), and *w/o BDA* and *w/o MiA* ($p = .022$). Turning to SA, the main effect of method was also significant ($\chi^2(4) = 39.947$, $p < .001$, $W = .499$). Post-hoc tests showed that *Ours* and *matrix* ($p = .022$), and *Ours* and *w/o BDA* ($p < .001$) conditions were significantly different. Other differences were found in the following pairs: *matrix* and *w/o MiA* ($p = .019$), *linear* and *w/o BDA* ($p = .015$), and *w/o BDA* and *w/o MiA* ($p = .012$). For SS, a significant main effect was found for *method* ($\chi^2(4) = 22.827$, $p < .001$, $W = .285$). Post-hoc analysis only revealed significant differences between *Ours* and *matrix* ($p = .015$), *linear* ($p = .041$), and *w/o BDA* ($p = .022$). No other significant differences were observed for VEQ+, SA, and SS.

VHPQ: We utilized a subscale ABP (Appearance and Behavior Plausibility) of VHPQ, and its internal consistency was satisfied with the accepted level ($\alpha_{VHPQ} = .945$). A significant main effect of *method* was found ($\chi^2(4) = 45.522$, $p < .001$, $W = .569$), and post-hoc tests showed significant differences in the pairs of *Ours* and *matrix* ($p = .006$), *Ours* and *w/o BDA* ($p < .001$). Significant contrasts were also identified for *matrix* and *linear* ($p = .008$), *matrix* and *w/o MiA* ($p = .005$), *linear* and *w/o BDA* ($p = .016$), and *w/o BDA* and *w/o MiA* ($p = .009$). All other pairs were not significantly different.

Facial Animation Realism: Lastly, the Real scale demonstrated excellent internal consistency ($\alpha_{Real} = .963$). We found a significant main effect of *method* ($\chi^2(4) = 37.130$, $p < .001$, $W = .464$). Pairwise comparisons revealed significant differences between *Ours* and *matrix* ($p = .018$), and *Ours* and *w/o BDA* ($p = .001$). Other pairs that showed significant differences were *linear* and *w/o BDA* ($p = .017$), and *w/o BDA* and *w/o MiA* ($p = .028$). No other pairs showed significant differences on Real.

6 Discussion

Our experimental results indicate that OFERA consistently outperforms the baselines (*matrix*, *linear*) and ablation models (*w/o BDA*, *w/o MiA*) in both objective expression mapping accuracy and subjective user assessments. In the quantitative and qualitative evaluations of expression mapping, OFERA achieved lower parameter and vertex errors than the *matrix*- and *linear*-based baselines, with particularly notable improvements in highly dynamic facial regions such as the eyebrows and mouth. These results indicate that transforming blendshape signals into a controllable expression parameter space benefits from non-linear modeling, and support the design choice of adopting an MLP-based Expression Parameter Mapper (EPM). In particular, the reduced vertex-wise reconstruction errors and clearer geometric structures observed in mesh-level visualizations suggest that the proposed non-linear mapping better captures the underlying geometric and structural characteristics of facial expressions than linear alternatives. Building on these objective improvements in expression mapping, we further examine how such gains translate into perceived realism, controllability, and embodiment in immersive VR environments through a user study.

These advantages were further reflected in the user study. In the statistical analysis, our method achieved significantly higher scores in virtual embodiment, self-identification, plausibility, and facial animation realism compared to *matrix*-based mapping and *w/o BDA* ablation conditions. Compared to the more competitive models (*linear* and *w/o MiA*), *Ours* reduced distortions more effectively than *linear* and preserved richer expressiveness than *w/o MiA*, leading to higher overall embodiment and realism. Since *w/o MiA* represents an ablation setting, the differences between *w/o MiA* and *Ours* were expected to be smaller, yet *Ours* still showed an advantage in expressiveness and controllability.

These trends were confirmed by participant feedback. *Ours* was described as *"most natural and human-like"*(P2), *"best followed my face and expressions"*(P8), and *"accurately reflected even subtle muscle movements"*(P5). In contrast, *linear* was noted for detail but still *"heavily distorted"*(P11), while the *w/o MiA* model was stable but *"less expressive"*(P13). As a result, these observations indicate that *Ours* achieves better performance in terms of stability and expressiveness. Finally, our system also demonstrated practical usability in VR, even under headset occlusion. Unlike baseline conditions that failed to reproduce eye movements (*"the eyes did not move at all"*, P18), *Ours* could reproduce such facial dynamics. Overall, these findings demonstrate that *Ours* provides a solid foundation for realistic, controllable virtual avatar expressions, while also delivering greater embodiment and self-identification to users than alternative mapping methods.

Beyond mapping accuracy and perceptual realism, our findings highlight the importance of system-level integration for practical avatar deployment in VR. Rather than introducing a new avatar representation, OFERA demonstrates that combining headset-available sensing, calibrated parameter mapping, and lightweight real-time communication can effectively bridge the gap between occluded user input and expressive photorealistic avatars. This suggests that realistic avatar control in VR can be achieved through careful coordination across sensing, representation, and rendering stages, even under strict latency and access constraints. Such a system-oriented perspective complements recent advances in avatar modeling and is particularly relevant for deployable VR telepresence applications.

6.1 Limitations and Future Work

The proposed system involves a stack of constraints arising from multiple stages of representation and parameter transformation, which can limit the expressiveness of the final avatar despite explicit mitigation efforts. In particular, OFERA relies on headset-provided blendshape signals as the sole input modality, inherently constraining expressiveness to the capacity of the blendshape space and limiting the accurate reproduction of subtle facial motions such as micro-expressions. These constraints are further influenced by differences in blendshape definitions and amplitude calibration across VR headsets, as well as discrepancies between the Meta Quest Pro blendshape pipeline and the commercial mesh avatar used for pseudo paired data construction. Although Blendshape Distribution Alignment (BDA) and Mapper-integrated Avatar (MiA) are introduced to reduce distribution mismatch and training–inference inconsistency across stages, residual constraints may still propagate through the pipeline. Moreover, the fidelity of facial deformations ultimately depends on the representation capacity of the underlying Gaussian avatar backbone model [67], and certain expression patterns may not be fully captured if the backbone is insufficiently trained. At the system level, to ensure real-time performance, OFERA fixes spherical harmonics and opacity during runtime, which limits support for dynamic relighting or appearance variations. Finally, our user study focused on observing overall user perceptions of the proposed system and therefore did not explicitly investigate other factors related to avatar representation, such as age, individual facial characteristics (e.g., face shape and expression style), or environmental lighting.

Future work will explore richer or more unified expression representations that reduce stacked pipeline constraints while preserving real-time performance, as well as more efficient system designs that enable dynamic appearance attributes. We also plan to extend the framework to support cross-headset calibration and evaluate robustness across heterogeneous devices, along with broader user studies conducted under more diverse participant profiles and usage conditions.

7 Conclusion

We propose **OFERA** (Occluded Facial Expression to Realistic Avatar), a real-time system that reconstructs occluded facial expressions of VR headset users as photorealistic 3D Gaussian avatars. OFERA is composed of three modules: (i) the *Blendshape Distribution Alignment* (BDA), which adapts headset-specific blendshapes to a canonical input space suitable for inference; (ii) the *Expression Parameter Mapper* (EPM), which maps the aligned blendshape signals into an expression parameter space for controlling Gaussian head avatars; and (iii) the *Mapper-integrated Avatar* (MiA), which ensures that the Gaussian avatar is trained to follow the output distribution of the EPM. By integrating these modules, OFERA establishes an end-to-end pipeline that takes blendshape data captured from a VR headset and renders a Gaussian avatar with realistic expressions in real-time within the VR environment. Our experiments demonstrated that the proposed EPM outperforms baseline models in both quantitative and qualitative evaluations, and user studies further validated the effectiveness of the BDA and MiA modules. This confirms that OFERA enables both realistic appearance and natural real-time expression control, which are critical for immersive telepresence in VR.

Acknowledgments

This work was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. CRC21015); 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 supported by the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2026-RS-2022-00156435) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1] S. Bai, T.-L. Wang, C. Li, A. Venkatesh, T. Simon, C. Cao, G. Schwartz, R. Wrench, J. Saragih, Y. Sheikh, et al. Universal facial encoding of codec avatars from vr headsets. *arXiv preprint arXiv:2407.13038*, 2024. 3

[2] J. N. Bailenson, A. C. Beall, J. Loomis, J. Blascovich, and M. Turk. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators & Virtual Environments*, 13(4):428–441, 2004. 1, 2

[3] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, 15(4):359–372, 2006. 1, 2

[4] S. Baker, J. Waycott, R. Carrasco, R. M. Kelly, A. J. Jones, J. Lilley, B. Dow, F. Batchelor, T. Hoang, and F. Vetere. Avatar-mediated communication in social vr: an in-depth exploration of older adult interaction in an emerging communication platform. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–13, 2021. 1, 2

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 157–164. ACM Press/Addison-Wesley Publishing Co., 2023. 2

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2

[7] T. Chen, Y. Li, S. Tao, H. Lim, M. Sakashita, R. Zhang, F. Guimbretiere, and C. Zhang. Neckface: Continuously tracking full facial expressions on neck-mounted wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–31, 2021. 3

[8] T. Chen, B. Steeper, K. Alsheikh, S. Tao, F. Guimbretière, and C. Zhang. C-face: Continuously reconstructing facial expressions by deep learning contours of the face with ear-mounted miniature cameras. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*, pp. 112–125, 2020. 3

[9] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–9, 2024. 3

[10] T. Combe, R. Fribourg, L. Detto, and J.-M. Normand. Exploring the influence of virtual avatar heads in mixed reality on social presence, performance and user experience in collaborative tasks. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2206–2216, 2024. 1, 2

[11] R. Daněček, M. J. Black, and T. Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022. 2

[12] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019. 2

[13] H.-B. Duan, M. Wang, J.-C. Shi, X.-C. Chen, and Y.-P. Cao. Bakedavatar: Baking neural fields for real-time head avatar synthesis. *ACM Transactions on Graphics (ToG)*, 42(6):1–17, 2023. 3

[14] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2, 3, 4, 6

[15] 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. 2

[16] 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. 6

[17] A. D. Fraser, I. Branson, R. C. Hollett, C. P. Speelman, and S. L. Rogers. Expressiveness of real-time motion captured avatars influences perceived animation realism and perceived quality of social interaction in virtual reality. *Frontiers in Virtual Reality*, 3:981400, 2022. 6

[18] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8649–8658, 2021. 2, 3

[19] S. Giebenhain, T. Kirschstein, M. Rünz, L. Agapito, and M. Nießner. Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024. 3

[20] 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. 6

[21] P.-W. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, and J. Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18653–18664, 2022. 2, 3

[22] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5784–5794, 2021. 2, 3

[23] Q. He, X. Ji, Y. Gong, Y. Lu, Z. Diao, L. Huang, Y. Yao, S. Zhu, Z. Ma, S. Xu, et al. Emotalk3d: High-fidelity free-view synthesis of emotional 3d talking head. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024. 3

[24] S. Kang, A. Nguyen, B. Yoon, K. Kim, and W. Woo. Gender differences in perceiving avatar face and interpersonal distance: Exploring realism and social presence in mixed reality. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 101–110. IEEE, 2024. 2, 6

[25] S. Kang, H. Song, B. Yoon, K. Kim, and W. Woo. The influence of emotion-based prioritized facial expressions on social presence in avatar-mediated remote communication. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1147–1156. IEEE, 2024. 2, 6

[26] S. Kang, B. Yoon, K. Kim, J. Gratch, and W. Woo. How collaboration context and personality traits shape the social norms of human-to-avatar identity representation. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[27] 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. 1, 2, 3, 5

[28] D. Kim and C. Harrison. Pantœnna: Mouth pose estimation for ar/vr headsets using low-profile antenna and impedance characteristic sensing. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–12, 2023. 3

[29] T. Kirschstein, S. Qian, S. Giebenhain, T. Walter, and M. Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 5

[30] P. Kullmann, T. Schell, T. Menzel, M. Botsch, and M. E. Latoschik. Coverage of facial expressions and its effects on avatar embodiment, self-identification, and uncanniness. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 6

[31] S. U. Lee, J. Kim, and J. Lee. Effects of reward schedule and avatar visibility on joint agency during vr collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4372–4382, 2023. 1, 2

[32] J. Li, C. Cao, G. Schwartz, R. Khirodkar, C. Richardt, T. Simon, Y. Sheikh, and S. Saito. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024. 3

[33] K. Li, R. Zhang, S. Chen, B. Chen, M. Sakashita, F. Guimbretière, and C. Zhang. Eyeecho: continuous and low-power facial expression tracking on glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024. 2, 3

[34] L. Li, Y. Li, Y. Weng, Y. Zheng, and K. Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10747–10757, 2025. 3

[35] 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. 2, 3, 4, 5, 6

[36] Y. Li, T. Zhang, X. Zeng, Y. Wang, H. Zhang, and Y. Chen. Auglasses: Continuous action unit based facial reconstruction with low-power imus on smart glasses. *arXiv preprint arXiv:2405.13289*, 2024. 2, 3

[37] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou, X. Zhe, N. Iwamoto, B. Zheng, and M. J. Black. 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. 1144–1154, 2024. 3, 5

[38] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 3

[39] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 4, 5

[40] S. Ma, Y. Weng, T. Shao, and K. Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–10, 2024. 3

[41] 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. 6

[42] J. Martinez, E. Kim, J. Romero, T. Bagautdinov, S. Saito, S.-I. Yu, S. Anderson, M. Zollhöfer, T.-L. Wang, S. Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 1, 5

[43] 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. 2, 3

[44] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016. 2, 3

[45] C. Patel, S. Bai, T.-L. Wang, J. Saragih, and S.-E. Wei. Fast registration of photorealistic avatars for vr facial animation. In *European Conference on Computer Vision*, pp. 407–423. Springer, 2024. 2, 3

[46] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20299–20309, 2024. 2, 3

[47] 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. 6

[48] 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. 3

[49] M. Slater, B. Spanlang, M. V. Sanchez-Vives, and O. Blanke. First person experience of body transfer in virtual reality. *PloS one*, 5(5):e10564, 2010. 1, 2

[50] 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. 3

[51] H. Song, S. Yang, and W. Woo. Fast texture transfer for xr avatars via barycentric uv conversion. *arXiv preprint arXiv:2508.19518*, 2025. 2

[52] 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. 2

[53] L. Song, P. Liu, L. Chen, G. Yin, and C. Xu. Tri 2-plane: Thinking head avatar via feature pyramid. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024. 3

[54] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv preprint arXiv:1610.03151*, 2016. 2, 3

[55] 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. 3

[56] C. Wang, D. Kang, H. Sun, S. Qian, Z. Wang, L. Bao, and S.-H. Zhang. Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26274–26284, 2025. 3

[57] J. Wang, J.-C. Xie, X. Li, F. Xu, C.-M. Pun, and H. Gao. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2, 3

[58] S.-E. Wei, J. Saragih, T. Simon, A. W. Harley, S. Lombardi, M. Perdoch, A. Hypes, D. Wang, H. Badino, and Y. Sheikh. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (ToG)*, 38(4):1–16, 2019. 2, 3

[59] F. Weidner, G. Boettcher, S. A. Arboleda, C. Diao, L. Sinani, C. Kunert, C. Gerhardt, W. Broll, and A. Raake. 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. 6

[60] Y. Wu, V. Kakaraparthi, Z. Li, T. Pham, J. Liu, and P. Nguyen. Bioface-3d: Continuous 3d facial reconstruction through lightweight single-ear biosensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pp. 350–363, 2021. 2, 3

[61] Y. Xu, L. Wang, X. Zhao, H. Zhang, and Y. Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–10, 2023. 3

[62] P. Yan, R. Ward, Q. Tang, and S. Du. Gaussian déjà-vu: Creating controllable 3d gaussian head-avatars with enhanced generalization and personalization abilities. *arXiv preprint arXiv:2409.16147*, 2024. 3, 5

[63] X. Yao, C. Yu, L. Hu, Y. Jin, Y. Gao, and Z. Jin. Imuface: Real-time, low-power, continuous 3d facial reconstruction through earphones. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pp. 614–615, 2025. 2, 3

[64] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, 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

[65] B. Yoon, H.-i. Kim, S. Y. Oh, and W. Woo. Evaluating remote virtual hands models on social presence in hand-based 3d remote collaboration. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 520–532. IEEE, 2020. 2

[66] B. Yoon, J.-e. Shin, H.-i. Kim, S. Y. Oh, D. Kim, and W. Woo. Effects of avatar transparency on social presence in task-centric mixed reality remote collaboration. *IEEE transactions on visualization and computer graphics*, 29(11):4578–4588, 2023. 2

[67] 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. 2, 3, 5, 9

[68] H. Zhu, H. Yang, L. Guo, Y. Zhang, Y. Wang, M. Huang, M. Wu, Q. Shen, R. Yang, and X. Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):14528–14545, 2023. 2

[69] W. Zielonka, T. Bolkart, and J. Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pp. 250–269. Springer, 2022. 2, 4, 5

[70] W. Zielonka, T. Bolkart, and J. Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4574–4584, 2023. 3, 5