PCHMR: Empowering and Benchmarking Human Mesh Recovery in Privacy-Constrained Real-World Settings

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning directly on user-side data is an effective way to improve the performance of widely adopted data-driven HMR models. However, doing so for existing HMR methods often assumes aggregating user-side data from real-world HMR-deployed devices to a central training server, posing a significant privacy risk as sensitive human images are transmitted. How can human mesh recovery be evaluated and improved in privacy-constrained real-world settings? This paper serves as a benchmark study of this problem. We conduct a comprehensive benchmark in which state-of-the-art HMR models are trained under federated and secure-aggregation variants that avoid raw-image centralization under explicit threat-model assumptions and benchmark their performance under a wide array of realistic client-scale and data-heterogeneity settings. We document that common HMR training pipelines are built around centralized data access and quantify how representative HMR backbones behave when that assumption is removed. Furthermore, to study the data bottleneck of privacy-constrained HMR training, we propose a local annotation and fine-tuning pipeline enhanced with depth foundation models, with which collaboratively trained HMR models can be locally tailored to the end user's distribution. We demonstrate its effectiveness with results on in-the-wild data while clarifying that the reported personalization gains are measured against DePoser-generated pseudo-ground truth. This benchmark aims to support future work on privacy-constrained HMR models and their real-world deployment and evaluation.

1 Introduction

Advancements in human mesh recovery (HMR) methods Lin et al. (2021b); Cho et al. (2022); Dou et al. (2023); Li et al. (2023a); Lin et al. (2021c); Li et al. (2021) have enabled a wide range of real-world applications, such as virtual reality Apple Inc. (2024); Meta Platforms, Inc. (2024a); VRChat (2024), augmented reality Meta Platforms, Inc. (2024b), human behavior analysis Liu* et al. (2022); Guo* et al. (2023); Zhang et al. (2023; 2024), autonomous driving NotaTeslaApp (2023) and virtual try-on FXGear (2024). Accurate human mesh recovery is always desired for those human-centric applications.

For general learning-based applications like medical image computing Guan et al. (2024); Adnan et al. (2022); Rehman et al. (2023) or on-device speech recognition Brydinskyi et al. (2024); Azam et al. (2023a;b), fine-tuning the model *directly on user-side data* is one of the most effective ways of enhancing the model performance by minimizing the domain gap Mairittha et al. (2020); Lee et al. (2023). To our best knowledge, existing HMR frameworks Kanazawa et al. (2018); Lin et al. (2021b); Dou et al. (2023); Cho et al. (2022); Choi et al. (2020); Li et al. (2021; 2023b;a); Lin et al. (2023) are typically developed and evaluated under the setting where all training data are on a single device, which we refer to as the *centralized* setting. As a result, HMR model training using existing frameworks usually assumes the collection and aggregation of sensitive data from user devices to a central training server. This raises privacy and compliance concerns, which is increasingly challenging, if not infeasible, under strict privacy regulations such as the General Data Protection Regulation (GDPR) European Parliament and Council of the European Union (2016). A naïve raw-data-local solution would be training the HMR model locally on each edge device, leveraging the scarce collection of user data. However, since the mapping from a monocular image to a 3D human's pose and

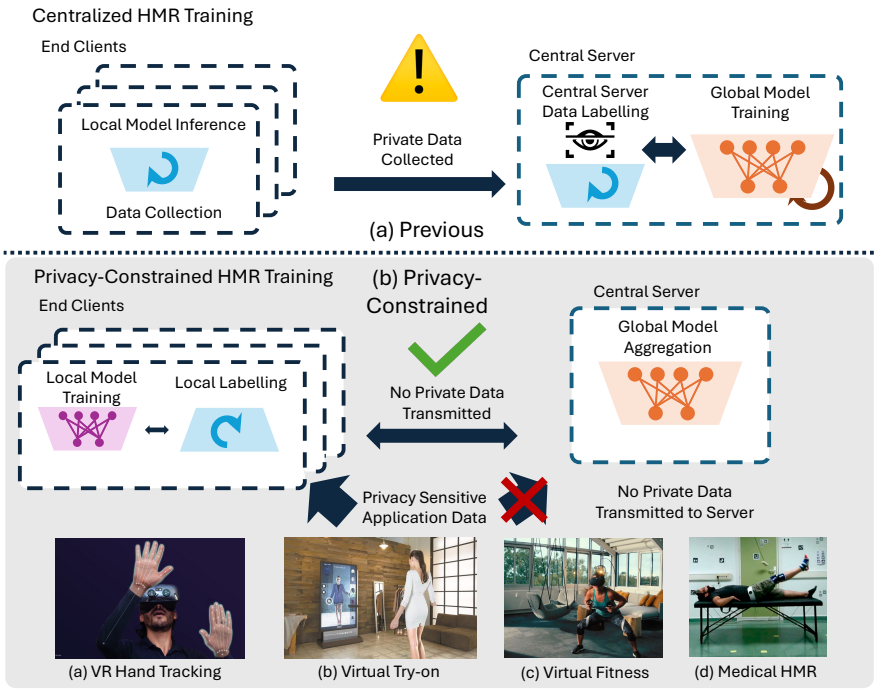


Figure 1: Comparison between centralized and privacy-constrained HMR frameworks. Existing HMR frameworks commonly assume centralized training (a), which requires uploading sensitive user data and therefore raises privacy concerns. In comparison, PCHMR benchmarks privacy-constrained settings for HMR models (b), where raw data remains local and only model-side information is exchanged under the specified collaborative-training protocol. This opens a wide range of applications with additional supervision from locally learned annotation and allows improvement of HMR model performance with *privacy-constrained real-world human data*.

shape is highly complex, training such a model with limited real-world data collected from only a single edge device would result in significantly inferior performance compared to centralized training; See Appendix. F. This poses a *performance-privacy dilemma*: the desire to effectively utilize user data on the edge to enhance the model's performance, conflicting with the principle of avoiding raw-data centralization.

Despite tremendous efforts in HMR Kanazawa et al. (2018); Lin et al. (2021b); Dou et al. (2023); Cho et al. (2022); Choi et al. (2020); Li et al. (2021; 2023b;a); Lin et al. (2023); Fiche et al. (2025) where models are centrally trained and evaluated under ideal conditions, we argue that a practical and critical question in the application of HMR remains unexplored: *How does state-of-the-art methods perform under privacy-constrained real-world settings, where centralized aggregation of data from user devices is infeasible due to privacy concerns?*

This work presents a systematic benchmark study of HMR models Dou et al. (2023); Cho et al. (2022)[1] under a comprehensive series of privacy-constrained, real-world settings, answering the aforementioned question. In all of our experiments, the training is conducted with a client-server model, where each client holds a part of the training data, which is never shared. Under this framework, we then scale up to a large number of clients, where we found that the HMR models converge with minimal accuracy loss even with 1,000 clients. In addition, since each HMR-equipped device user could have a different usage pattern, we simulate real-world data non-uniformity with three schemes of varying complexity. When data heterogeneity is pronounced, collaboratively trained HMR models can match, and in some settings slightly exceed, their centrally trained counterparts. We interpret this as evidence that federated training need not degrade HMR accuracy under the tested protocols; whether the occasional gains arise from aggregation-induced regularization, training-budget effects, or variance in centralized baselines remains an open question. Furthermore, we evaluate a secure aggregation variant with MPC Knott et al. (2021). This variant protects the aggregation channel

[1]Although we select two representative transformer-based methods, the framework discussed in this paper is agnostic to the choice of HMR architecture.

under an honest-but-curious server assumption and prevents the server from directly observing individual client updates, but it should not be read as an attack-based privacy evaluation of the released global HMR model. With less than 2x the run time in our small-scale measurement, MPC appears feasible enough to merit further system-level study.

In privacy-constrained HMR, raw images cannot be transmitted to a central server, so locally annotated data becomes essential both for collaborative training and model personalization. However, user- and environment-specific scenarios vary widely, so the annotation pipeline must be especially robust, which makes optimization-based annotators appealing due to their stability and accuracy. Unlike existing methods Bogo et al. (2016); Pavlakos et al. (2019a; 2018b; 2019b); Xu et al. (2019) that rely on sparse cues such as 2D keypoints or masks, whose performance can be limited by detector quality and monocular ambiguity, we introduce *DePoser*, a depth-aware HMR annotation pipeline.

DePoser leverages the large human vision foundation model Sapiens Khirodkar et al. (2024), a 2B-parameter model trained on 300M+ in-the-wild images, to provide body-part relative depth cues, and then aligns these predictions to metric scale via a lightweight linear regressor trained with a general absolute depth estimator, preserving Sapiens' structure. At the same time, we exploit Sapiens' accurate and robust 2D joint predictions to provide stronger optimization signals. This yields 2.5D keypoints that couple precise localization with scale-invariant depth, reducing ambiguity and improving the annotator in the evaluated settings. DePoser is fully monocular, deployable on common devices from a single RGB image input. We validate our pipeline against Pavlakos et al. (2019a) and benchmark downstream HMR models on an internet-sourced VR dataset annotated by DePoser. Because the current comparison changes both the 2D keypoint source and the depth term, we treat the improvement as evidence for the combined DePoser pipeline rather than as a component-level ablation of depth alone.

We show that locally fine-tuning models on such DePoser-generated annotations enhances HMR performance in privacy-constrained collaborative training setups, enabling on-device models to self-collect pseudo-supervision. In summary, our contributions are:

- We identify and evaluate the centralized-data assumption in state-of-the-art HMR training pipelines, which is essential for real-world applications of such models.
- We provide a comprehensive study of state-of-the-art HMR models, investigating their performance under privacy-constrained training settings of various complexity and assessing the feasibility of collaborative training for HMR models.
- Local annotators are essential for privacy-constrained HMR training. Motivated by the depth ambiguity issue of previous optimization-based HMR annotators, we further propose a local HMR annotation pipeline enhanced by depth foundation models, which can be integrated with the proposed collaborative training framework while making the pseudo-label dependency explicit.

2 Related Work

There has been a plethora of recent works on 3D representations of the human body or body parts, such as 3D skeleton Martinez et al. (2017); Mehta et al. (2017); Yang et al. (2018b), 3D whole-body mesh Li et al. (2021; 2023a); Dou et al. (2023); Cho et al. (2022); Choi et al. (2020); Lin et al. (2021b), and 3D hand/face recovery Wu et al. (2024); Shimada et al. (2023); Moon et al. (2024). Due to the large amount of prior work, we focus on the 3D whole-body mesh recovery in the following section for their highest relevance.

2.1 Human Mesh Recovery

Optimization-based Human Mesh Recovery. This approach estimates the 3D mesh that is consistent with the 2D observation under various constraints by directly optimizing with standard body model parameters. After the seminal work SMPLify Bogo et al. (2016), different objective functions Pavlakos et al. (2018b); Xu et al. (2019); Pavlakos et al. (2019b) are proposed to analyze the predicted 3D mesh that is compatible with the 2D image observation. In addition, factors such as shape, pose, and segmentation provide additional

priors to the optimization problem, such as HuMoR Rempe et al. (2021), Pose-NDF Tiwari et al. (2022) and others Güler et al. (2018); Ke et al. (2022). This type of method has also been expanded to regress detailed human body, hand, and face at the same time Pavlakos et al. (2019a); Zioulis & O'Brien (2023); Lin et al. (2024). Optimization-based methods apply gradient-based optimization with a neural network and perform iteratively to obtain accurate mesh results such as HUND Zanfir et al. (2021) and LVD Corona et al. (2022). This method generally yields a higher quality of mesh compared with the regression-based approach; however, due to their iterative optimizing nature, these methods are often incapable of real-time inference.

Regression-based Human Mesh Recovery. This method generates 3D human mesh directly from pixels with an end-to-end network. Various methods use intermediate representations such as joints Xue et al. (2021), segmentation Rueegg et al. (2020), UV maps Yao et al. (2019) or heatmaps Moon & Lee (2020) to supply additional information for the mesh regression. In addition, forward kinematics and inverse kinematics may be considered at the same time as additional constraints for the neural network, such as HybrIK Li et al. (2021), HybrIK-X Li et al. (2023b) and NIKI Li et al. (2023a). This representation is then passed into a decoder network, the output of which could be classified into the following two categories: *parametric output* and *non-parametric output*. The first is in the parameter space of parametric models such as Loper et al. (2023), while the latter is in 3D space, such as joints and mesh vertex coordinates. SOTA performance has been achieved by transformer-based methods such as METRO Lin et al. (2021b), MeshGraphormer Lin et al. (2021c) and FastMETRO Cho et al. (2022), which are non-parametric approaches excelling at modeling long-range dependency, making them robust even under occlusions. Transformer efficiency is optimized with token pruning methods such as in TORE Dou et al. (2023).

2.2 Federated Learning

Federated Learning (FL) McMahan et al. (2017) is a family of raw-data-local distributed machine learning paradigms that differs from the ordinary training paradigm (we refer to it as *Centralized Training*): FL does not require the raw training data to be sent back to the central server to train the model. Instead, it trains a local model and sends the local model updates to the central server to *aggregate* a global model. This design reduces raw-data exposure and can be combined with secure aggregation or differential privacy, but FL alone does not constitute a complete privacy guarantee against all leakage channels, including information that may be inferable from model updates or a released global model Kairouz et al. (2019).

FL has been successfully applied to diverse fields—recommender systems Du et al. (2021); Caballero et al. (2016); Duan et al. (2020); Muhammad et al. (2020), finance Cheng et al. (2019), edge computing Ye et al. (2020); Jiang et al. (2020); Wang et al. (2021); Liu et al. (2021), industrial applications Tang et al. (2022); Zhao et al. (2020); Li et al. (2020a), and healthcare Ouyang et al. (2021); Dayan et al. (2021); Ma et al. (2022); Chen et al. (2020)—thanks to its collaborative and raw-data-local nature. FL also demonstrates strong performance in computer vision Liu et al. (2020); Li et al. (2019), natural language processing Yang et al. (2018a); Lin et al. (2021a), and automatic speech recognition Azam et al. (2023b); Zhu et al. (2022). However, to our knowledge, it has not yet been used to address data grounding challenges in real-world monocular HMR. Existing FL-based computer vision solutions mostly cover image classification Liu et al. (2020) or segmentation Li et al. (2019), with the only human-related FL work focusing on pose estimation from either images or point clouds Zhuang et al. (2024); Wang et al. (2024)—a task that differs significantly from monocular HMR and yields only basic 2D poses. In contrast, our approach directly regresses the 3D human mesh from standard RGB images, offering a more versatile and practical solution without the need for specialized hardware.

Other distributed training paradigms are also relevant. Split learning trains different model segments across client and server and exchanges intermediate activations rather than complete model updates Vepakomma et al. (2018), while FedSplit studies operator-splitting based federated optimization Pathak & Wainwright (2020). These approaches may be useful for future HMR systems with different communication, compute, and leakage trade-offs. In this paper, we focus on federated parameter sharing because it directly tests whether existing HMR backbones can be trained without centralizing raw user images, and because it is readily supported by mature benchmark infrastructure.

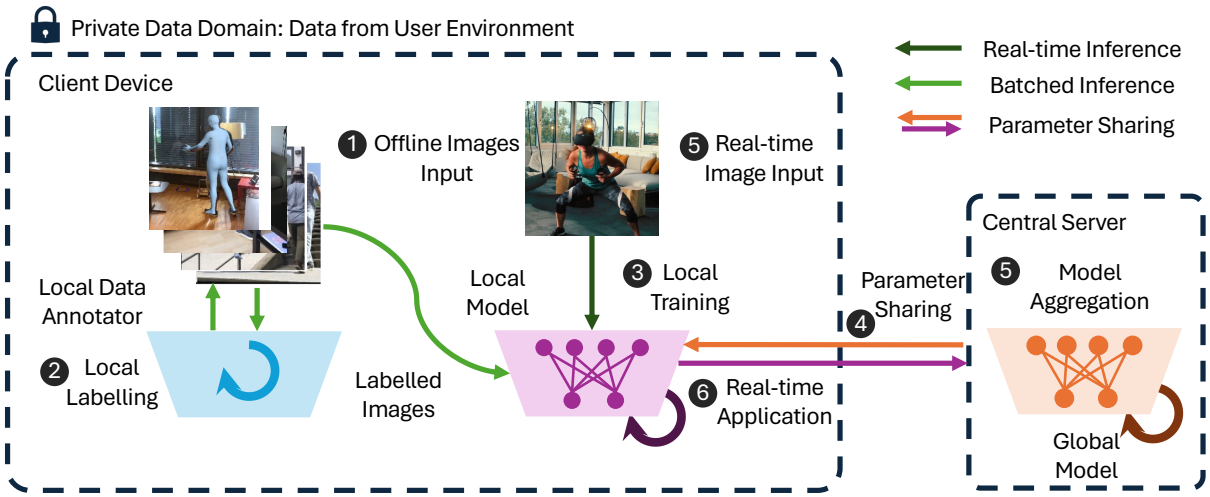


Figure 2: Overview of the Pipeline for Privacy-Constrained HMR. This shows an overview of the real-world pipeline for privacy-constrained HMR training on a client device. Data labeling can be done offline via a local data annotator (1,2). After that, the label of user-generated data is used to collaboratively train a new global model with privacy constraints (3, 4). Note that during the process the user data never leaves its private domain. The global model is then used to provide better local data representation for local real-time inference (5,6). PCHMR benchmarks implementations of such privacy-constrainted pipeline under various real-world settings.

3 Method

Central to the PCHMR framework is an HMR model $\mathcal{M}$, with parameters w takes a monocular image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ as input and outputs a 3D human mesh $\mathcal{V} \in \mathbb{R}^{6890\times 3}$ in the SMPL Loper et al. (2023) topology.

Under our privacy-constrained setting, the HMR model is trained on a distributed HMR dataset, divided into disjoint subsets $\mathcal{D} = \cup_{k=1}^{n}\mathcal{D}_k$ and $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset\ \forall i \neq j$, where each subset $\mathcal{D}_k$ is locally collected and annotated on device k. Each local dataset consists of n_k locally annotated image-human mesh pairs, $\mathcal{D}_k = \{(\mathcal{I}_i, \mathcal{V}_i)\})_{i=1}^{n_k}$. An overview of the PCHMR framework is shown in figure 2.

We first conduct a comprehensive benchmark of HMR in privacy-constrained real-world settings in section 3.1. Then, in Sec. 3.2, we introduce and evaluate a local data annotator, named DePoser, which enables human mesh annotation for privacy-constrained HMR training.

3.1 Privacy-Constrained Benchmark Construction

We create a client-server framework to train HMR models in a raw-data-local way to construct our benchmark. Specifically, we use the Flower Beutel et al. (2020) library to simulate a federated training setup. While Secure Multi-Party Computation (MPC) Knott et al. (2021) and Differential Privacy (DP) Dwork (2006); Abadi et al. (2016) are also considered for privacy-preserving machine learning methods, due to the computation and communication overhead of completely setting up MPC training and the performance degradation of DP under large scale models (such as SOTA HMR models), we primarily focus on benchmarking under the Federated Learning McMahan et al. (2017) approach. We consider the basic setup for federated training and the secure-aggregation variant enabled by MPC. Additional details on MPC and DP are discussed in Appendix. section E.

Privacy Scope and Threat Model. Our benchmark evaluates utility when raw images and mesh annotations remain on client devices. In vanilla FL, the server still receives individual model updates, so the protocol reduces raw-data centralization but does not by itself rule out update-level or released-model leakage. In the secure-aggregation variant, we assume an honest-but-curious server and use MPC to prevent the server from observing individual client updates before aggregation. We do not claim formal differential privacy and do not perform membership-inference, gradient-inversion, or model-inversion attacks in this work; these attack-based privacy evaluations are important future extensions.

We conduct a series of comprehensive experiments on two popular HMR models Cho et al. (2022); Dou et al. (2023) on various settings that closely reflect the complexities of the real-world training environment:

Large Number of Participating Devices. In a typical real-world privacy-respecting setup, a large number of user devices are often involved Huba et al. (2022); Bonawitz et al. (2019). In addition, due to varied usage patterns, it is impossible for each user to participate in every global fitting round. Based on these observations, we perform experiments with 10, 100, and 1000 simulated user devices (clients), with data randomly partitioned for each device to simulate locally collected and stored private training data. The number of our clients is limited by the total training dataset size (∼500,000 samples), averaging about hundreds to a few thousand data points per client, depending on the setup. In each training round, 10 clients are randomly selected as participants, which trains one epoch using the locally partitioned data. Then, the model weights are uploaded to the central server and aggregated, while the local dataset remains local for raw-data locality.

Heterogeneous Data Distribution. In the privacy-constrained HMR training setup, the training data are collected and annotated locally. Thus, each user's locally annotated dataset can differ substantially in data distribution and volume from another user's local dataset due to variation in human identity, body shape, attire, and usage environment. We employ multiple methods to simulate this highly non-uniform data distribution:

1) We use Latent Dirichlet Allocation Blei et al. (2003) (LDA) with the parameter α to control the degree of heterogeneity of the distribution. See Appendix. H.1 for a visualization.

2) We create a mixed dataset comprising a series of datasets Ionescu et al. (2013); Lassner et al. (2017); Lin et al. (2014); Andriluka et al. (2014); Mehta et al. (2017) with 3DPW von Marcard et al. (2018) and assign clients according to datasets. This creates a more heterogeneous setup, benchmarking the SOTA HMR model's capability under challenging training settings.

3) To simulate real-world fine-tuning, where a centrally pre-trained HMR model is fine-tuned with user-side data, we perform privacy-constrained fine-tuning on the in-the-wild 3DPW Von Marcard et al. (2018) dataset. We further adopt the challenging setup of assigning each scene to a participating client, which resembles real-world use cases where each client uses their edge device in a different environment.

Privacy-Constrained Personalized Fine-tuning. Data distribution alignment is essential in real-world deployment of HMR models, which could be effectively achieved by directly fine-tuning on the end user's data distribution. To evaluate the effectiveness of model personalized fine-tuning, where federated-trained HMR model is further fine-tuned on each user's local dataset, we collect a dataset of over 3000 images from the internet and in-house VR usage footage, comprising 4 different scenes with users using VR and AR devices. The internet subset is drawn from publicly accessible videos, and the in-house subset is recorded with participant consent for research use. We use this dataset as a personalization stress test rather than as a demographic coverage benchmark. We further randomly partition the dataset into training and validation sets, fine-tuning the model for 2 epochs on the training sets and then evaluating on the validation sets.

3.2 DePoser: Depth-Aware Local Annotator

For many HMR applications, fine-tuning on user-side data is the most effective way of aligning with the target user's distribution. However, privacy constraints and the sensitive nature of human images make the central aggregation of such images impossible. As a result, an accurate and local annotation pipeline is central to the privacy-constrained training of HMR models.

Optimization-based HMR methods Bogo et al. (2016); Pavlakos et al. (2019a;b); Xu et al. (2019) are suitable candidates for local annotation due to their robustness and accuracy. These methods often use 2D keypoints to guide the optimization process of human mesh. However, relying solely on 2D keypoints introduces monocular ambiguity, often leading to solutions that produce plausible 2D re-projections but suffer from poor 3D accuracy. Inspired by SMPLify-X Pavlakos et al. (2019a) and DICE Wu et al. (2024), we devise a depth-aware monocular local annotation pipeline named DePoser, where the mesh is supervised not only by 2D keypoints, but also affine-invariant depth information in the z-direction. First, we extract depth map and 2D keypoints from a large human foundational model, Sapiens Khirodkar et al. (2024). by taking advantage of

a model pre-trained on vast in-the-wild human data, our model benefits from a strong human prior, enabling more reliable downstream mesh prediction with 2.5D depth cues for in-the-wild user-side scenarios. Second, our method does not require expensive hardware such as RGB-D sensors, LiDAR sensors, or synchronized multi-view cameras, making it possible to deploy DePoser to ordinary consumer-grade devices with only monocular cameras.

Specifically, given an RGB image, we first use Sapiens-pose Khirodkar et al. (2024) to extract 2D human keypoints. Then, the corresponding pseudo-depth of each keypoint is extracted from a depth map obtained using Sapiens-depth Khirodkar et al. (2024). Subsequently, we minimize the following objective function with the L-BFGS optimizer Nocedal & Wright (2006):

$$E(\beta, \theta) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{C} E_\mathcal{C} + \lambda_d E_d, \tag{1}$$

where E_J is the 2D joints projection loss, E_{θ_b} is the pose prior Pavlakos et al. (2019a) loss, $E_\alpha = \sum_{i \in (\text{elbows,knees})} \exp(\theta_i)$ is the elbow/knee bending regularizer, $E_\beta = \|\beta\|^2$ is the shape regularizer, E_C the interpenetration penalty, and E_d the SILog Eigen et al. (2014) loss:

$$E_d = \frac{1}{N} \sum_{i \in \text{joints}} d_i^2 - \frac{1}{2} \left(\frac{1}{N} \sum_{i \in \text{joints}} d_i^2 \right)^{1/2}, \tag{2}$$

where $d_i = \log z_i - \log \hat{z}_i$ and $\hat{z}_i, z_i$ are the predicted and pseudo-GT depth value of the i-th joint.

4 Evaluation

4.1 Experiments Setup

Datasets. Following previous works Zhang et al. (2019); Cho et al. (2022); Dou et al. (2023); Lin et al. (2021b), we use the Human3.6M Ionescu et al. (2013), 3DPW von Marcard et al. (2018), UP-3D Lassner et al. (2017), COCO Lin et al. (2014), MPII Andriluka et al. (2014), and MuCo-3DHP Mehta et al. (2017) datasets to train our models. We evaluate our model using Human3.6M Ionescu et al. (2013), 3DPW von Marcard et al. (2018) datasets.

In addition, we also curate a dataset from internet video footage and in-house VR usage. This dataset consists of over 3000 image-mesh pairs with 4 different scenes and identities. We conduct local fine-tuning and testing experiments on this dataset. For the internet-sourced subset, we use publicly accessible videos and retain source metadata for reproducibility. For the in-house subset, participants consented to research use of the recordings. Since the mesh labels are generated by DePoser rather than captured from an independent multi-view or marker-based system, we report these experiments as pseudo-labeled personalization studies and do not claim independent ground-truth coverage or demographic representativeness. Annotation quality is controlled through the quantitative DePoser comparison on 3DPW in Sec. C and by excluding cases where the local fitting pipeline fails to produce a valid body fit.

Evaluation Metrics. Following previous works Zhang et al. (2019); Cho et al. (2022); Dou et al. (2023); Lin et al. (2021b), We use *Mean Per Joint Position Error* (MPJPE) Ionescu et al. (2013); Zhou et al. (2018) and *Procrustes Aligned Mean Per Joint Position Error* (PA-MPJPE) Pavlakos et al. (2018a) to evaluate our models' accuracy. *Mean Per Vertex Position Error* (MPVPE) Pavlakos et al. (2018a) are also calculated when ground-truth 3D mesh is available. All the error metrics are in millimeters (mm).

Simulation. Following previous works that adopted the federated learning paradigm Jithish et al. (2023); Lubana et al. (2022); Madill et al. (2022); Gao et al. (2022); Jiang et al. (2023), we conduct our experiments in a simulated environment using Flower Beutel et al. (2020) to avoid dealing with low-level implementations such as communication and resource scheduling.

Reproducibility and Release Plan. All reported client partitions are generated with fixed random seeds and fixed partition metadata. Upon publication, we will release the Flower-based benchmark harness, evaluation scripts, hyperparameter files, deterministic client-splitting scripts, and the partition indices used

in our experiments. We will also release the DePoser implementation, including the metric-scale depth alignment component. For data with redistribution restrictions, we will release the permitted annotations, source metadata, split files, and scripts needed to reconstruct the benchmark when users have access to the underlying images. We will not redistribute raw in-house frames unless the corresponding consent and license terms allow it; for internet-sourced videos, we will release source URLs, timestamps, split files, and a takedown/update policy rather than republishing media whose license is uncertain.

4.2 Privacy-Constrained Evaluations

We choose two representative SOTA HMR models Cho et al. (2022); Dou et al. (2023) for their robustness and accuracy. Note that, due to the well-modularized nature of the HMR community, other models can be easily incorporated into our benchmark pipeline. For the chosen models, we performed extensive experiments in a series of simulated settings that mimic scenarios encountered in real-world privacy-constrained training. We group the experiments into categories and outline each category in the following sections.

#Clients (Total / Per Round)	Model	MPJPE	PA-MPJPE	Raw Data Local?
Centralized	FM-S	57.98	40.62	✗
10 / 10	FM-S	56.85	40.52	✓
100 / 10	FM-S	59.31	41.48	✓
1000 / 10	FM-S	60.82	44.53	✓
Centralized	TORE-S	63.88	41.99	✗
10 / 10	TORE-S	61.27	41.60	✓
100 / 10	TORE-S	62.92	43.11	✓
1000 / 10	TORE-S	66.04	44.04	✓

Table 1: Effect of Number of Clients. This table shows the HMR results (MPJPE and PA-MPJPE) for the PCHMR baseline experiments (`10/10`), the Independent and Identically Distributed (IID) data 100 clients experiments (`100/10`) and IID data 1000 clients experiments (`1000/10`) with their best evaluation results. All experiments select 10 clients per round uniformly at random. For comparison, we also add the centralized baseline trained with the same model and the same dataset.

Optimizer	FM-S Model		TORE-S Model	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
IID Distribution				
FedAvg	59.31	41.48	62.92	43.11
FedOpt	58.49	41.43	62.82	42.34
FedProx	57.71	41.26	63.41	43.17
LDA $\alpha = 0.5$ Distribution				
FedAvg	57.80	39.82	58.91	40.94
FedOpt	55.41	39.36	59.34	41.92
FedProx	56.46	39.58	61.57	41.10
LDA $\alpha = 1$ Distribution				
FedAvg	59.63	39.80	62.39	41.76
FedOpt	58.10	40.10	59.57	42.10
FedProx	56.45	39.28	58.96	41.97
LDA $\alpha = 10$ Distribution				
FedAvg	56.09	39.84	60.96	42.25
FedOpt	58.25	40.47	58.83	41.29
FedProx	57.33	40.27	58.96	41.57

Table 2: Effect of Federated Optimizer Performance on HMR model with different federated optimizer (FedAvg, FedOpt, FedProx) on IID data and LDA skewed data. The results are measured with 100 clients experiments (`100/10`) with 10 clients sampled per round.

4.2.1 Performance on Large Number of Clients

We evaluate the performance of HMR models when trained under a simulated Federated Learning environment with 10, 100, and 1000 participating devices, where 10 devices are randomly selected to participate at each round. The environment has a client-server configuration, where each client trains the model for one epoch and sends the model weights to a central server for aggregation using the FedAvg McMahan et al. (2017) algorithm.

As shown in Tab. 1, both FastMETRO Cho et al. (2022) and TORE Dou et al. (2023) trained under the FedAvg optimizer maintain competitive accuracy even in the most challenging setting of 1000 clients, with roughly a 2–4 mm drop in MPJPE and PA-MPJPE. In the 10-client setting, the HMR models even exhibit numerically higher accuracy than the centrally-trained counterparts. This underscores the robustness of the benchmarked state-of-the-art HMR models under the complex federated setting, even in the most challenging settings of 1000 clients. Qualitative results for HMR models trained with the 100-client configuration are shown in Fig. 4.

While achieving a performance close to centrally-trained models, one caveat is that the HMR models train much more slowly: they require many more rounds to converge, with the most complex 1000-client setting requiring 6000 rounds. Please refer to Appendix. A for more details.

4.2.2 Performance on Skewed Data Distribution

In the real world, user data from user devices may vary in data distribution and sample size due to differences in identity, environment, and habits. Therefore, we manually skew the data distribution in various configurations to simulate this phenomenon and benchmark the performance of HMR models to assess their robustness under skewed data distribution. We investigate three types of heterogeneous data distributions: **Latent Dirichlet Allocation (LDA)**, **Mixed Dataset**, and **Natural Partition Data**. More details on the experiment setup are in Appendix section B.

Latent Dirichlet Allocation. As shown in figure 3, the models under a high heterogeneous LDA setup perform the best under the most skewed distribution, $\alpha = 0.1$, with both MPJPE and PA-MPJPE error lower than the centralized baseline for both configurations.

A plausible explanation is that local updates from highly non-uniform clients introduce an aggregation-induced regularization effect akin to meta learning Nichol et al. (2018), but our current experiments do not isolate this hypothesis from centralized-run variance or remaining training-budget differences. In addition, across a wide range of α, the model performance is consistently competitive compared to the centralized baseline. These results support the robustness of the benchmarked HMR models Cho et al. (2022); Dou et al. (2023) under highly heterogeneous setups close to real-world scenarios.

Mixed Dataset. As shown in Tab. 3, the mixed dataset result demonstrates that the HMR model incorporating mixed-source training data performs similarly well under the majority dataset (Human3.6M), sometimes even achieving a performance gain of 1–2mm. This suggests that HMR models trained on the

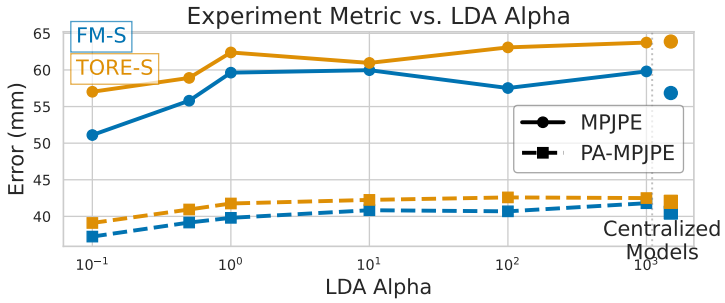


Figure 3: Data Distribution Evaluation. Performance metrics for heterogenous data distribution settings with different LDA α, evaluated with the Human3.6M Ionescu et al. (2013) dataset.

Setting	Model	MPJPE	PA-MPJPE	Raw Data Local?
Centralized	FM-S	61.38	41.63	✗
PCHMR	FM-S	58.28	40.42	✓
Centralized	TORE-S	63.89	41.40	✗
PCHMR	TORE-S	61.59	41.94	✓

Table 3: Mixed Dataset Training Comparison between FastMETRO-S Cho et al. (2022) and TORE-S Dou et al. (2023) trained under centralized or privacy-constrained settings on a mixed dataset composed of Human3.6M, UP-3D, COCO, MPI-INF-3DHP, and MPII Ionescu et al. (2013); Lassner et al. (2017); Lin et al. (2014); Mehta et al. (2017); Andriluka et al. (2014), evaluated on Human3.6M.

Setting	Model	MPJPE	PA-MPJPE	MPVPE	Raw Data Local?
Centralized	FM-S	84.92	54.69	97.60	✗
PCHMR	FM-S	84.66	54.78	97.42	✓
PCHMR (Natural Part.)	FM-S	86.83	55.61	99.66	✓
Centralized	TORE-S	87.97	55.35	101.88	✗
PCHMR	TORE-S	87.55	54.08	101.86	✓
PCHMR (Natural Part.)	TORE-S	88.00	55.07	102.50	✓

Table 4: Performance of PCHMR on 3DPW This table shows the reconstruction accuracy when PCHMR is fine-tuned on the 3DPW dataset. We further evaluate the model's performance on the highly challenging natural scene partition (Natural Part.), where each scene is assigned to a device. PCHMR reaches comparable performance on 3DPW fine-tuning compared with the centralized setting. PCHMR is trained with 100 devices and 10 randomly selected each round, while the natural partition is 20 devices with 5 randomly selected each round.

privacy-constrained federated training framework can achieve competitive performance on the overall data distribution compared to the traditional centralized training framework.

Natural Partition Data. As shown in Tab. 4, under both partition schemes, FastMETRO and TORE show similar performance with their centralized counterpart, with only 1 mm of accuracy difference. The results show that privacy-constrained distributed fine-tuning of HMR models can achieve similar performance compared to centralized fine-tuning, even under the challenging natural partition scheme.

Dataset	Before/After	Model	MPVPE	MPJPE	PA-MPJPE
VR-runner	Before PCHMR	FM-S	109.50	114.51	60.07
VR-runner	After PCHMR	FM-S	93.23	58.24	43.16
VR-game-1	Before PCHMR	FM-S	176.24	172.57	78.21
VR-game-1	After PCHMR	FM-S	65.57	62.30	43.82
VR-game-2	Before PCHMR	FM-S	150.64	161.82	84.36
VR-game-2	After PCHMR	FM-S	82.53	69.49	53.57
Oculus	Before PCHMR	FM-S	103.92	102.22	61.62
Oculus	After PCHMR	FM-S	56.76	57.92	39.94

Table 5: Training accuracy before and after local data fine-tuning introduced in PCHMR. The metrics are calculated on the locally annotated test data as described in Sec. C.

4.2.3 Comparison between Federated Optimizer

In the previous section, experiment results confirmed the robustness of FastMETRO Cho et al. (2022) and TORE Dou et al. (2023) under different forms of data skewness. We now benchmark the HMR models on three popular federated learning algorithms, namely FedAvg McMahan et al. (2017), FedOpt Asad et al. (2020) and FedProx Li et al. (2020b). These two more sophisticated federated optimizers offer better adaptiveness and robustness than FedAvg.

As shown in Tab. 2, in most of the settings evaluated, the three optimizers provided similarly in terms of both MPJPE and PA-MPJPE, with no clear trend of superiority and inferiority. This suggests the robustness of

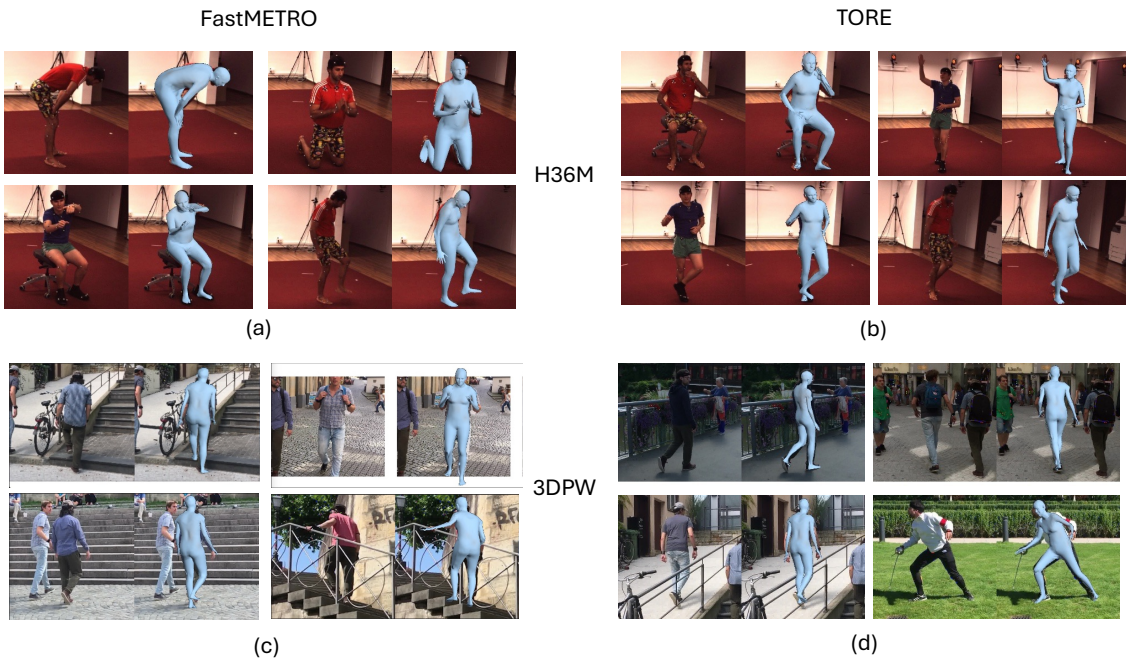


Figure 4: Qualitative results of PCHMR on Human3.6M Ionescu et al. (2013) (a,b) and 3DPW Von Marcard et al. (2018) (c,d), trained with 100 clients and 10 randomly sampled per round. The backbone is FastMETRO Cho et al. (2022) (a,c) and TORE Dou et al. (2023) (b,d).

HMR models coupled with the simplest FedAvg McMahan et al. (2017) under a highly skewed data setting (LDA $\alpha = 0.5$).

4.2.4 Personalized Finetuning

Although HMR models could maintain competitive performance under challenging real-world multi-device collaborative training settings, further personalized fine-tuning is often desired to tailor the model to a specific user's data distribution. Therefore, we benchmark further fine-tuning a collaboratively pre-trained FastMETRO Cho et al. (2022) model on additional in-the-wild datasets that reflect real-world use-cases, such as VR gaming and sports. We use our local data annotator, DePoser, to annotate pseudo-ground-truths for the collected dataset. For local fine-tuning, we divide each scene of our dataset into an 80/20 random split and fine-tune the pre-trained model for two epochs on this dataset.

As shown in Fig. 5, our personalized fine-tuned model reconstructs human meshes that more accurately resemble the human pose and shape after fine-tuning on the Pseudo GT. Quantitatively, as shown in Tab. 5, significant improvements up to 100 mm in MPVPE are observed after local fine-tuning, with the most improvement observed in VR-game-1, a challenging home VR gaming setting. Because the evaluation labels in this table are generated by DePoser, these numbers measure consistency with the local pseudo-label distribution rather than accuracy against independent ground truth. In addition, we perform the personalized fine-tuning using the 3DPW dataset, which exhibits similar performance improvement. Details are shown in Appendix. I. This demonstrates the strong prior of an HMR model pre-trained on a large amount of data could achieve effective personalization with only two epochs of fine-tuning, significantly improving end-user performance through model personalization.

4.3 Local Annotation Pipeline

We now compare DePoser vs. SMPLify-X Pavlakos et al. (2019a) to demonstrate the effect of the incorporation of the Sapiens Khirodkar et al. (2024) pose and depth priors. This comparison evaluates the full DePoser pipeline; it does not disentangle how much of the gain comes from replacing the keypoint source, adding the depth term, or using metric-scale depth alignment. The accuracy of the local annotation pipeline

Model	MPVPE	MPJPE	PA-MPJPE
SMPLify-X Pavlakos et al. (2019a)	198.66	181.07	87.17
DePoser (Ours)	**162.60**	**151.19**	**71.14**

Table 6: Comparison between DePoser and SMPLify-X Pavlakos et al. (2019a) on sub-sampled 3DPW Von Marcard et al. (2018) dataset. DePoser demonstrates improved reconstruction accuracy compared to SMPLify-X under the combined Sapiens pose and depth prior.

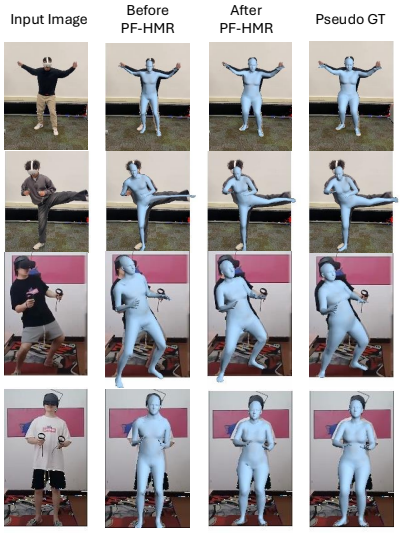


Figure 5: Qualitative comparison of FastMETRO Cho et al. (2022) before and after PCHMR fine-tuning on a locally-annotated dataset, pseudo-ground truth (GT) generated by DePoser.

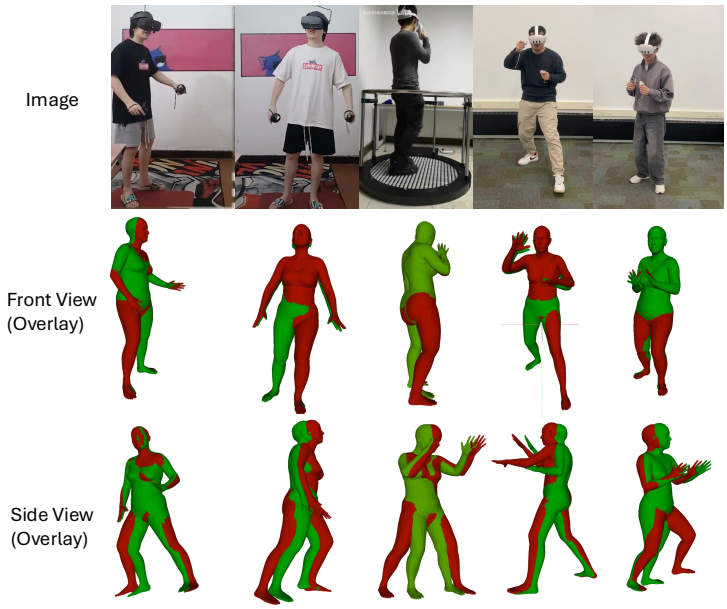


Figure 6: Comparison between our DePoser (green) vs. SMPLify-X (red) Pavlakos et al. (2019a). DePoser exhibits more accurate foot placing and hand positioning under the combined Sapiens pose and depth prior; component attribution requires future ablation.

is critical in enabling HMR models to perform privacy-constrained training. More details are available at Appendix section C.

Quantitative Analysis. We compare the human reconstruction accuracy of our DePoser and SMPLify-X Pavlakos et al. (2019a) on a 300-image randomly sub-sampled set of the in-the-wild dataset 3DPW von Marcard et al. (2018), which resembles complex user-side conditions. As shown in Tab. 6, DePoser demonstrates improved accuracy compared to SMPLify-X, with more than 30 mm accuracy improvement in MPVPE and 15 mm improvement in PA-MPJPE, supporting the effectiveness of the combined local annotation pipeline under complex conditions.

Qualitative Analysis. As shown in Fig. 6, although the front-view results of DePoser and SMPLify-X Pavlakos et al. (2019a) are similar, the additional depth supervision in DePoser regresses more accurate foot and arm positionings in the z-axis (depth) compared to SMPLify-X, which is evident in the side view. This visual evidence is consistent with the quantitative improvement, while the exact contribution of depth relative to the updated pose prior requires future component-level ablation.

5 Limitations and Future Works

HMR-aware Aggregation Strategy. Motivated by the data distribution challenges, our LDA results suggest that aggregation under heterogeneous client distributions may sometimes behave like a regularizer for HMR. This remains a hypothesis rather than a demonstrated mechanism, and future work should use

matched-budget controlled studies to separate aggregation effects from optimization variance and schedule differences.

Device-specific and Personalized Learning. In real-world human mesh recovery, each device would introduce unique data perturbation into the client data distribution, such as smearing, tearing, and fuzziness. However, there is no existing solution for simulating this issue. Having a personalized model adapted to device perturbations would be a promising direction for future work.

Privacy Evaluation Scope. Our experiments focus on HMR utility when raw images remain local and on the runtime cost of a secure-aggregation variant. We do not perform membership-inference, gradient-inversion, or model-inversion attacks, and therefore do not claim that the released global models are private against such attacks. Attack-based privacy evaluation and stronger formal guarantees are important future work.

Local Annotation Bias. DePoser produces SMPL-parametric pseudo-annotations, which are then used to supervise non-parametric HMR regressors. The personalized fine-tuning gains should therefore be interpreted as improved alignment to DePoser's pseudo-label distribution, not as independent ground-truth validation. The current DePoser comparison also evaluates the combined Sapiens-pose, Sapiens-depth, and metric-alignment pipeline rather than isolating each component.

6 Conclusion

In this work, we introduced PCHMR, a benchmark for evaluating human mesh recovery (HMR) models under privacy-constrained, raw-data-local training settings. Through extensive experiments across various complex and realistic scenarios, we show that state-of-the-art HMR models can be collaboratively trained with a large number of clients and heterogeneous data while incurring only minimal accuracy loss. At the same time, our study reveals remaining challenges, including slower convergence and degraded performance when generalizing to out-of-distribution data, highlighting important gaps between centralized and privacy-constrained training regimes. To address the data availability bottleneck in privacy-constrained HMR, we proposed *DePoser*, a local data annotation pipeline that conditions not only on 2D keypoints but also on pseudo-depth information. Comparisons against SMPLify-X demonstrate that DePoser improves pseudo-annotation quality under the combined Sapiens pose and depth priors. Using DePoser, we curated and annotated an in-the-wild dataset of VR use-cases, and showed that privacy-constrained fine-tuning on this dataset yields substantial gains for HMR model personalization on end-user devices when evaluated against the same pseudo-label source. We hope that our benchmark, analysis, and the DePoser pipeline will serve as a foundation for future research on privacy-constrained human mesh recovery, including more robust collaborative training algorithms, stronger privacy guarantees, and better on-device personalization strategies.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Apple Inc. Apple Vision Pro. `https://www.apple.com/apple-vision-pro/`, 2024. Accessed: 2024-11-13.

Muhammad Asad, Ahmed Moustafa, and Takayuki Ito. Fedopt: Towards communication efficiency and privacy preservation in federated learning. *Applied Sciences*, 10(8):2864, 2020.

Sheikh Shams Azam, Tatiana Likhomanenko, Martin Pelikan, et al. Importance of smoothness induced by optimizers in fl4asr: Towards understanding federated learning for end-to-end asr. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023a.

Sheikh Shams Azam, Martin Pelikan, Vitaly Feldman, Kunal Talwar, Jan Silovsky, and Tatiana Likhomanenko. Federated learning for speech recognition: Revisiting current trends towards large-scale asr. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*, 2023b.

Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 561–578. Springer, 2016.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

Vitalii Brydinskyi, Dmytro Sabodashko, Yuriy Khoma, Michal Podpora, Alexander Konovalov, and Volodymyr Khoma. Enhancing automatic speech recognition with personalized models: Improving accuracy through individualized fine-tuning. *IEEE Access*, 2024.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36:41727–41764, 2023.

Speedtest by Ookla. United states median country speeds, 2024. URL `https://www.speedtest.net/global-index/united-states#fixed`. Accessed: 2024-11-14.

Alberto Caballero, Teresa García-Valverde, Fernando Pereñíguez-Garcia, and Juan A. Botía Blaya. Activity recommendation in intelligent campus environments based on the eduroam federation. *J. Ambient Intell. Smart Environ.*, 8:35–46, 2016. URL `https://api.semanticscholar.org/CorpusID:41402920`.

Shaoqi Chen, Dongyu Xue, Guohui Chuai, Qiang Yang, and Qi Liu. Fl-qsar: a federated learning based qsar prototype for collaborative drug discovery. *bioRxiv*, 2020. URL `https://api.semanticscholar.org/CorpusID:214362498`.

Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, and Qiang Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 36:87–98, 2019. URL `https://api.semanticscholar.org/CorpusID:59292020`.

Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pp. 342–359. Springer, 2022.

Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pp. 769–787. Springer, 2020.

Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *European Conference on Computer Vision*, pp. 146–165. Springer, 2022.

Ittai Dayan, Holger R. Roth, Aoxiao Zhong, Ahmed El Harouni, Amilcare Gentili, Anas Zainul Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J. Wood, Chien-Sung Tsai, Chih-Hung Wang, Chun-Nan Hsu, C K Lee, Peiying Ruan, Daguang Xu, Dufan Wu, Eddie Huang, Felipe Campos Kitamura, Griffin Lacey, Gustavo César de Antônio Corradi, Gustavo Nino, Hao-Hsin Shin, Hirofumi Obinata, Hui Ren, Jason C. Crane, Jesse Tetreault, Jiahui Guan, John W. Garrett, Joshua D. Kaggie, Jung Gil Park, Keith Dreyer, Krishna Juluru, Kristopher Kersten, Marcio Aloisio Bezerra Cavalcanti Rockenbach, Marius George Linguraru, Masoom A. Haider, Meena Abdelmaseeh, Nicola Rieke, Pablo F. Damasceno, Pedro Mário Cruz e Silva, Pochuan Wang, Sheng Xu, Shuichi Kawano, Sira Sriswasdi, Soo Young Park, Thomas M. Grist, Varun Buch, Watsamon Jantarabenjakul, Weichung Wang, Won Young Tak, Xiang Li, Xihong Lin, Young Joon Kwon, Abood Quraini, Andrew Feng, Andrew N. Priest, Baris I Turkbey, Benjamin Scott Glicksberg, Bernardo Canedo Bizzo, Byung Seok Kim, Carlos Tor-Díez, Chia-Cheng Lee, Chia-Jung Hsu, Chin Lin, Chiu-Ling Lai, Christopher P. Hess, Colin B. Compas, Deepeksha Bhatia, Eric Karl Oermann, Evan Leibovitz, Hisashi Sasaki, Hitoshi Mori, Isaac Yang, Jae Ho Sohn, Krishna Nand Keshava Murthy, Lijuan Fu, Matheus Ribeiro Furtado de Mendonça, Michael Fralick, Min Kyu Kang, Mohammad Adil, Natalie Gangai, Peerapon Vateekul, Pierre Elnajjar, Sarah E Hickman, Sharmila Majumdar, Shelley L. McLeod, Sheridan L. Reed, Stefan Gräf, Stephanie A. Harmon, Tatsuya Kodama, Thanyawee Puthanakit, Tony Mazzulli, Vitor Lima de Lavor, Yothin Rakvongthai, Yu Rim Lee, Yuhong Wen, Fiona J. Gilbert, Mona G. Flores, and Quanzheng Li. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature Medicine*, 27:1735 – 1743, 2021. URL `https://api.semanticscholar.org/CorpusID:237536154`.

Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15143–15155, 2023.

Yongjie Du, Deyun Zhou, Yu Xie, Jiao Shi, and Maoguo Gong. Federated matrix factorization for privacy-preserving recommender systems. *Appl. Soft Comput.*, 111:107700, 2021. URL `https://api.semanticscholar.org/CorpusID:237666062`.

Sijing Duan, Deyu Zhang, Yanbo Wang, Lingxiang Li, and Yaoxue Zhang. Jointrec: A deep-learning-based joint cloud video recommendation framework for mobile iot. *IEEE Internet of Things Journal*, 7:1655–1666, 2020. URL `https://api.semanticscholar.org/CorpusID:208111241`.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union, 2016. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504` (accessed 2023-03-29).

Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Mega: Masked generative autoencoder for human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5366–5378, 2025.

FXGear. Vr fashion. `http://www.fxgear.net/vr-fashion`, 2024. Accessed: 2024-05-17.

Yan Gao, Titouan Parcollet, Salah Zaiem, Javier Fernandez-Marques, Pedro PB de Gusmao, Daniel J Beutel, and Nicholas D Lane. End-to-end speech recognition from federated acoustic models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7227–7231. IEEE, 2022.

Lucas Grativol, Mathieu Léonardon, Guillaume Muller, Virginie Fresse, and Matthieu Arzel. Federated learning compression designed for lightweight communications. In *2023 30th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1–4. IEEE, 2023.

Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern Recognition*, pp. 110424, 2024.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.

Yong Guo*, Zhiyang Dou*, Nan Zhang, Xiyue Liu, Boni Su, Yuguo Li, and Yinping Zhang. Student close contact behavior and covid-19 transmission in china's classrooms. *PNAS Nexus*, 2(5):pgad142, 2023.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL `https://api.semanticscholar.org/CorpusID:206594692`.

Tzu-Ming Harry Hsu, Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *ArXiv*, abs/1909.06335, 2019a. URL `https://api.semanticscholar.org/CorpusID:202572978`.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019b.

Dzmitry Huba, John Nguyen, Kshitiz Malik, Ruiyu Zhu, Mike Rabbat, Ashkan Yousefpour, Carole-Jean Wu, Hongyuan Zhan, Pavel Ustinov, Harish Srinivas, et al. Papaya: Practical, private, and scalable federated learning. *Proceedings of Machine Learning and Systems*, 4:814–832, 2022.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Hui Jiang, Min Liu, Bo Yang, Qingxiang Liu, Jizhong Li, and Xiaobing Guo. Customized federated learning for accelerated edge computing with heterogeneous task targets. *Comput. Networks*, 183:107569, 2020. URL `https://api.semanticscholar.org/CorpusID:224979418`.

Lekang Jiang, Filip Svoboda, and Nicholas D Lane. Fdapt: Federated domain-adaptive pre-training for language models. *arXiv preprint arXiv:2307.06933*, 2023.

J Jithish, Bithin Alangot, Nagarajan Mahalingam, and Kiat Seng Yeo. Distributed anomaly detection in smart grids: a federated learning-based approach. *IEEE Access*, 11:7157–7179, 2023.

Peter Kairouz, H. B. McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary B. Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim Y. El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Oluwasanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, R. Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Xiaodong Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14:1–210, 2019. URL `https://api.semanticscholar.org/CorpusID:209202606`.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1140–1147, 2022.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pp. 206–228. Springer, 2024.

Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, and Laurens van der Maaten. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34:4961–4973, 2021.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6050–6059, 2017.

Kyungsu Lee, Haeyun Lee, Thiago Coutinho Cavalcanti, Sewoong Kim, Georges El Fakhri, Dong Hun Lee, Jonghye Woo, and Jae Youn Hwang. Fine-tuning network in federated learning for personalized skin diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 378–388. Springer, 2023.

Beibei Li, Yuhao Wu, Jiarui Song, Rongxing Lu, Tao Li, and Liang Zhao. Deepfed: Federated deep learning for intrusion detection in industrial cyber–physical systems. *IEEE Transactions on Industrial Informatics*, 17(8):5615–5624, 2020a.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3383–3393, 2021.

Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12933–12942, 2023a.

Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020b.

Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, Manuel Jorge Cardoso, and Andrew Feng. Privacy-preserving federated brain tumour segmentation. In *MLMI@MICCAI*, 2019. URL `https://api.semanticscholar.org/CorpusID:203627027`.

Bill Yuchen Lin, Chaoyang He, ZiHang Zeng, Hulin Wang, Yufen Huang, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. Fednlp: Benchmarking federated learning methods for natural language processing tasks. In *NAACL-HLT*, 2021a. URL `https://api.semanticscholar.org/CorpusID:238259354`.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1954–1963, 2021b.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12939–12948, 2021c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Hong Liu, Shuaipeng Zhang, Pengfei Zhang, Xinqiang Zhou, Xuebin Shao, Geguang Pu, and Yan Zhang. Blockchain and federated learning for collaborative intrusion detection in vehicular edge computing. *IEEE Transactions on Vehicular Technology*, 70:6073–6084, 2021. URL `https://api.semanticscholar.org/CorpusID:235589750`.

Hongbin Liu, Lun Wang, Om Thakkar, Abhradeep Thakurta, and Arun Narayanan. Differentially private parameter-efficient fine-tuning for large asr models. *arXiv preprint arXiv:2410.01948*, 2024.

Xiyue Liu*, Zhiyang Dou*, Lei Wang, Boni Su, Tianyi Jin, Yong Guo, Jianjian Wei, and Nan Zhang. Close contact behavior-based covid-19 transmission and interventions in a subway system. *Journal of Hazardous Materials*, 436:129233, 2022.

Yang Liu, Anbu Huang, Yu Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Hang Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. In *AAAI Conference on Artificial Intelligence*, 2020. URL `https://api.semanticscholar.org/CorpusID:210714181`.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. ACM, 2023.

Ekdeep Singh Lubana, Chi Ian Tang, Fahim Kawsar, Robert P Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering. *arXiv preprint arXiv:2205.11506*, 2022.

Zezhong Ma, Meng Zhang, Jiajia Liu, Aimin Yang, Hao Li, Jian Wang, Dianbo Hua, and Mingduo Li. An assisted diagnosis model for cancer patients based on federated learning. *Frontiers in Oncology*, 12, 2022. URL `https://api.semanticscholar.org/CorpusID:247193284`.

Evan Madill, Ben Nguyen, Carson K Leung, and Sara Rouhani. Scalesfl: a sharding solution for blockchain-based federated learning. In *Proceedings of the Fourth ACM International Symposium on Blockchain and Secure Critical Infrastructure*, pp. 95–106, 2022.

Nattaya Mairittha, Tittaya Mairittha, and Sozo Inoue. Improving activity data collection with on-device personalization using fine-tuning. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pp. 255–260, 2020.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2640–2649, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pp. 120–130. IEEE, 2018.

Meta Platforms, Inc. Meta Oculus. `https://www.meta.com/quest/`, 2024a. Accessed: 2024-11-13.

Meta Platforms, Inc. Project Aria Glasses. https://www.projectaria.com/, 2024b. Accessed: 2024-11-13.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *ArXiv*, abs/2008.03713, 2020. URL https://api.semanticscholar.org/CorpusID:221090782.

Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *Advances in Neural Information Processing Systems*, 36, 2024.

Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Z. Tragos, Barry Smyth, Neil J. Hurley, James Geraci, and A. Lawlor. Fedfast: Going beyond average for faster training of federated recommender systems. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. URL https://api.semanticscholar.org/CorpusID:221191466.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, pp. 270–302, 2006.

NotaTeslaApp. All Tesla Fsd Visualizations and What They Mean. https://www.notateslaapp.com/tesla-reference/636/all-tesla-fsd-visualizations-and-what-they-mean, 2023. Accessed: 2024-05-17.

NVIDIA Corporation. *NVIDIA Jetson AGX Orin Technical Brief*, 2021. URL https://www.nvidia.com/content/dam/en-zz/Solutions/gtcf21/jetson-orin/nvidia-jetson-agx-orin-technical-brief.pdf. Accessed: 2024-05-24.

Liwei Ouyang, Yong Yuan, Yumeng Cao, and Feiyue Wang. A novel framework of collaborative early warning for covid-19 based on blockchain and smart contracts. *Information Sciences*, 570:124 – 143, 2021. URL https://api.semanticscholar.org/CorpusID:233182297.

Reese Pathak and Martin J. Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7057–7066, 2020.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 459–468, 2018a. URL https://api.semanticscholar.org/CorpusID:13693947.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 459–468, 2018b.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael 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, 2019a.

Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 803–812, 2019b.

Muhammad Habib Ur Rehman, Walter Hugo Lopez Pinaya, Parashkev Nachev, James T Teo, Sebastin Ourselin, and M Jorge Cardoso. Federated learning for medical imaging radiology. *The British Journal of Radiology*, 96(1150):20220890, 2023.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499, 2021.

Nadine Rueegg, Christoph Lassner, Michael J. Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3d human pose and shape by cycling between representations. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:208278393.

D. Sculley. Web-scale k-means clustering. In *The Web Conference*, 2010. URL https://api.semanticscholar.org/CorpusID:6634147.

Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (ToG)*, 42(6):1–16, 2023.

Zhenheng Tang, Zhikai Hu, Shaohuai Shi, Yiu-ming Cheung, Yilun Jin, Zhenghang Ren, and Xiaowen Chu. Data resampling for federated learning with non-iid labels. In *Proceedings of the International Workshop on Federated and Transfer Learning for Data Sparsity and Confidentiality in Conjunction with IJCAI, Montreal, Canada*, pp. 21–22. FTLIJCAI, 2021.

Zhongyun Tang, Haiyang Hu, and Chonghuan Xu. A federated learning method for network intrusion detection. *Concurrency and Computation: Practice and Experience*, 34(10):e6812, 2022.

Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pp. 572–589. Springer, 2022.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://api.semanticscholar.org/CorpusID:5855042.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018.

Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 601–617, 2018.

VRChat. Vrchat. https://hello.vrchat.com/, 2024. Accessed: 2024-05-17.

Jiaxin Wang, Huanyu Deng, Yulong Wang, Jiexin Xie, Hui Zhang, Yang Li, and Shijie Guo. Multi-sensor fusion federated learning method of human posture recognition for dual-arm nursing robots. *Information Fusion*, 2024. URL https://api.semanticscholar.org/CorpusID:267947736.

Qizhao Wang, Qing Li, Kai Wang, Hong Wang, and Peng Zeng. Efficient federated learning for fault diagnosis in industrial cloud-edge computing. *Computing*, 103:2319 – 2337, 2021. URL https://api.semanticscholar.org/CorpusID:237072663.

Qingxuan Wu, Zhiyang Dou, Sirui Xu, Soshi Shimada, Chen Wang, Zhengming Yu, Yuan Liu, Cheng Lin, Zeyu Cao, Taku Komura, et al. Dice: End-to-end deformation capture of hand-face interactions from a single image. *arXiv preprint arXiv:2406.17988*, 2024.

Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7760–7770, 2019.

Lixia Xue, Meian Li, Liang Fan, Aixia Sun, and Tian Gao. Monocular vision ranging and camera focal length calibration. *Scientific Programming*, 2021(1):9979111, 2021.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *ArXiv*, abs/1812.02903, 2018a. URL https://api.semanticscholar.org/CorpusID:54457251.

Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5255–5264, 2018b.

Pengfei Yao, Zhen Fang, Fan Wu, Yao Feng, and Jiwei Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *ArXiv*, abs/1903.10153, 2019. URL https://api.semanticscholar.org/CorpusID:85498357.

Yunfan Ye, Shen Li, Fang Liu, Yonghao Tang, and Wanting Hu. Edgefed: Optimized federated learning based on edge computing. *IEEE Access*, 8:209191–209198, 2020. doi: 10.1109/ACCESS.2020.3038287.

Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14484–14493, 2021.

Nan Zhang, Xiyue Liu, Shuyi Gao, Boni Su, and Zhiyang Dou#. Popularization of high-speed railway reduces the infection risk via close contact route during journey. *Sustainable Cities and Society*, 99:104979, 2023.

Nan Zhang, Xueze Yang, Boni Su, and Zhiyang Dou. Analysis of sars-cov-2 transmission in a university classroom based on real human close contact behaviors. *Science of The Total Environment*, 917:170346, 2024.

Xiong Zhang, Qiang Li, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2354–2364, 2019. URL https://api.semanticscholar.org/CorpusID:67855669.

Ruijie Zhao, Yue Yin, Yong Shi, and Zhi Xue. Intelligent intrusion detection based on federated learning aided long short-term memory. *Physical Communication*, 42:101157, 2020.

Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):901–914, 2018.

Hanjing Zhu, Jindong Wang, Gaofeng Cheng, Pengyuan Zhang, and Yonghong Yan. Decoupled federated learning for asr with non-iid data. *ArXiv*, abs/2206.09102, 2022. URL https://api.semanticscholar.org/CorpusID:249889474.

Weiming Zhuang, Jian Xu, Chen Chen, Jingtao Li, and Lingjuan Lyu. Coala: A practical and vision-centric federated learning platform. *arXiv preprint arXiv:2407.16560*, 2024.

Nikolaos Zioulis and James F O'Brien. Kbody: Towards general, robust, and aligned monocular whole-body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6215–6225, 2023.

This supplementary material covers: More Experiment Details (section A); More Details on Skewed Data Distribution (section B); Local Data Annotation (section C); More Implementation Details (section D); Discussion on MPC and DP (section E); Discussion on Single-Device Training (section F); Discussion on Multiple Local Epoches (section G); Discussion on Performance of 3DPW Fine-tuning (section I); Discussion on Data Distribution (section H); Additional Experiments for On-Device Performance (section J).

A Experiment Hyperparameters

In Tab. 7 we present the hyperparameters for the experiments in Sec. 4 of the main paper.

Setting	**Model**	**Clients**	**Learning Rate**	**Epochs**
Centralized Baseline	FM-S	-	0.001	60
Human3.6M naïve on-device training	FM-S	-	0.001	60
3DPW Centralized Fine-tuning	FM-S	-	0.0001	20
PCHMR + IID Data 10 / 10	FM-S	10 / 10	0.001	200
PCHMR + IID Data 100 / 10	FM-S	100 / 10	0.001	2000
PCHMR + IID Data 1000 / 10	FM-S	1000 / 10	0.001	6000
PCHMR + LDA Distribution	FM-S	100 / 10	0.001	2000
PCHMR + CNN Cluster	FM-S	100 / 10	0.001	2000
PCHMR + 3DPW Finetuning	FM-S	100 / 10	0.0001	600
PCHMR + 3DPW Natural Part.	FM-S	20 / 5	0.0001	200
Centralized Baseline	TORE-S	-	0.001	60
Human3.6M naïve on-device training	TORE-S	-	0.001	60
3DPW Centralized Fine-tuning	TORE-S	-	0.0001	20
PCHMR + IID Data 10 / 10	TORE-S	10 / 10	0.0005	200
PCHMR + IID Data 100 / 10	TORE-S	100 / 10	0.0005	2000
PCHMR + IID Data 1000 / 10	TORE-S	1000 / 10	0.001	6000
PCHMR + LDA Distribution	TORE-S	100 / 10	0.0005	2000
PCHMR + CNN Cluster	TORE-S	100 / 10	0.0005	2000
PCHMR + 3DPW Finetuning	TORE-S	100 / 10	0.0001	600
PCHMR + 3DPW Natural Part.	TORE-S	20 / 5	0.0001	200

Table 7: The hyperparameters used by all PCHMR experiments. The two numbers in the Clients column denote the total number of clients and the number of clients randomly selected in each round.

B Skewed Data Distribution

We present the details of experiment setup for the three skewed data distributions discussed in section 4.2.2: **Latent Dirichlet Allocation (LDA)**, **Mixed Dataset**, and **Natural Partition Data**.

Latent Dirichlet Allocation. To test HMR model's performance under a heterogeneous data setting without any prior partition information available, we follow Tang et al. (2021); Hsu et al. (2019b); Grativol et al. (2023) to use Latent Dirichlet Allocation (LDA) Blei et al. (2003) for generating skewed training datasets to approximate real-world data distributions. The LDA partition scheme uses the α parameter to control the skewness of the generated partition with the label. A smaller α corresponds to a more skewed distribution, while a larger α indicates that the distribution is close to IID (independent and identically distributed). More detail for LDA is discussed in Appendix. H.1. For LDA, we investigate FastMETRO and TORE's performance with 100 clients and 10 clients per round, trained with datasets of varying skewness controlled by LDA.

Mixed Dataset. In addition to simulated label skewness, we also implemented a mixed dataset which blends the training datasets Ionescu et al. (2013); Lassner et al. (2017); Lin et al. (2014); Andriluka et al. (2014); Mehta et al. (2017). We allocate each client to a source dataset according to the dataset size ratio, which

Methods	Model	Round Time	Convergence
FL	TORE	139.8s	Converge
MPC	TORE	265.5s	Converge
DP	TORE	139.8s	Non-Converge

Table 8: Comparison of different privacy-preserving machine learning paradigms. The experiment uses the `10/10` TORE setting on an IID Human3.6M partition and reports average per-round runtime.

creates a heterogeneous training corpus across clients. The experiment is performed under the 100-client setting with 10 clients fitted per round.

Natural Partition Data. We also present a study on 3DPW Von Marcard et al. (2018), an in-the-wild human mesh dataset collected on various indoor and outdoor scenes, which provides a natural way to partition the dataset by scenes, resulting in a highly heterogeneous per-client data distribution. Under this data partitioning, we explore the performance of the HMR model trained using a privacy-constrained method. Further information of the 3DPW natural partition is available in Appendix. H.3. As a baseline, we first fine-tune pre-trained HMR models on the 3DPW dataset under the federated framework using IID partitioning scheme. Then, to explore the effect of this challenging partitioning scheme, we compare the IID-trained model to a model trained under the same dataset with natural partitioning. Qualitative results of the IID-finetuned model are shown in Fig.4.

C Local Data Annotation

Local data annotation, as described in Sec. 3.1 of the main paper, plays a critical role in improving the performance of PCHMR with client-side user data. Although there are many existing annotation approaches Bogo et al. (2016); Pavlakos et al. (2019a); Zioulis & O'Brien (2023); Lin et al. (2024), we choose a popular method SMPLify-X Pavlakos et al. (2019a) to serve as the backbone local data annotation. SMPLify-X is an optimization-based human mesh regressor, where a high-quality SMPL Loper et al. (2023) mesh of 6890 vertices can be obtained as the *pseudo-ground-truth mesh*. The input 2D keypoints and depth supervision information for DePoser optimization are obtained through Sapiens Khirodkar et al. (2024).

The model is pre-trained using the PCHMR framework for 2000 epochs with 100 clients and 10 clients per round under IID partition. We then fine-tune the model with the locally annotated data using a learning rate of 1×10^{-4} for 100 epochs, with other training settings such as optimizers consistent with the centralized baseline.

Remark. Notably, we currently employ an optimization-based approach to obtain pseudo-ground truth from monocular images. However, end devices possessing multi-view cameras could benefit from the additional 3D spatial information extracted by multiple cameras or depth sensors, thereby enhancing the local data annotation process using actual depth information with higher accuracy.

D Implementation Details

In our experiments, we develop and evaluate our model based on two backbones including FastMETRO Cho et al. (2022) and TORE Dou et al. (2023) using the Transformer encoder-decoder architecture. The details of this architecture are shown in Fig. 7. Please refer to the original papers for further details.

E Discussion on MPC and DP

We present a simple ablation setting comparing the MPC and DP regarding HMR model. For simplicity, we select the simplest `10/10` setting, i.e. partitioned data with 10 client, each time sampled 10 clients) with the TORE model using an IID partitioned Human3.6M training set. For MPC, we choose to use secure aggregation, which implements MPC at the aggregation stage and hides individual client updates from an

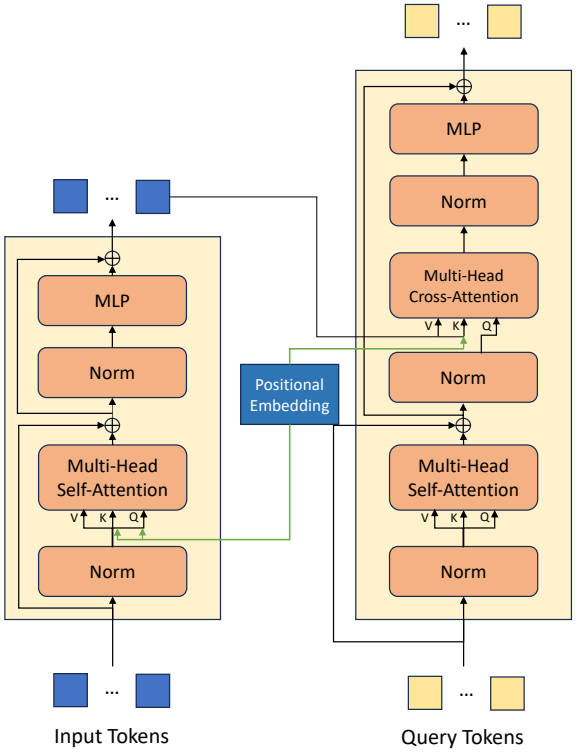


Figure 7: Details of the Transformer encoder-decoder architecture used in FastMETRO Cho et al. (2022) and TORE Dou et al. (2023).

honest-but-curious server before aggregation. This acts as a baseline setting to demonstrate the computation and communication overhead of the MPC protocol.

For DP, we choose to account with Renyi Differential Privacy Mironov (2017). We implement this with fixed clipping noise injection techniques. With the size of the Human3.6M dataset, we set $\delta = 1 \times 10^{-6}$ and calculate the respective ϵ privacy budget together with the noise multiplier set. Following suggestion on setting L2 Norm Bu et al. (2023), we set the L2 Norm to a small value, starting with 1×10^{-6}.

The results are presented in Tab 8. As demonstrated by the average round time, which is almost double that of FL, the MPC approach shows significant communication and computation overhead. However, the evaluation metric (MPJPE, PA-MPJPE) still shows similar performance with MPC, suggesting that MPC may be feasible to implement in HMR real-world settings with future improvement in hardware and communication. Even worse, although we attempted an extensive search on hyperparameter, we could not get our large HMR model to converge even with a small noise scale multiplier ($\eta = 1$), resulting in large $\epsilon = 199.7$ that offers a relatively small example-level DP guarantee. Due to the size of the SOTA model such as TORE, the curse of dimensionality Liu et al. (2024) makes the convergence hard, and how to pre-train a large-scale model with DP efficiently remains an open question for future research.

F Discussion on Single-Device Training

We investigate the performance of naïve on-device training compared with privacy-constrained collaborative training under PCHMR. We simulate single-device training with a small portion of the training dataset, and compare the accuracy with models trained with full data under centralized and privacy-constrained frameworks.

In Tab. 9, we pre-train the model using $5, 10, 20\%$ of the Human3.6M, which resulted in a large performance gap for both FastMETRO and TORE implementations. Specifically, the centralized training configuration

Settings	Model	MPJPE	PA-MPJPE	Raw Data Local?
Centralized Baseline	FM-S	56.85	40.52	✗
Naive 5% Data Training	FM-S	83.78	63.65	-
Naive 10% Data Training	FM-S	73.04	53.36	-
Naive 20% Data Training	FM-S	65.10	47.01	-
PCHMR 100% Data Training	FM-S	59.31	41.48	✓
Centralized Baseline	TORE-S	63.88	41.99	✗
Naive 5% Data Training	TORE-S	101.59	82.82	-
Naive 10% Data Training	TORE-S	84.82	63.52	-
Naive 20% Data Training	TORE-S	77.87	52.37	-
PCHMR 100% Data Training	TORE-S	62.92	43.11	✓

Table 9: Comparison with Naïve Single Device Training This table presents the performance metrics (MPJPE, PAMPJPE) for training using various sizes (5, 10, 20%) of the Human3.6M dataset to assess the on-device training scenario using only local data. Compared with naïve local training which suffers from poor performance due to data scarcity, our PCHMR could mitigate both issues due to the inter-device parameter sharing, achieving much higher performance while keeping raw data local. The PCHMR in this table is trained in a distributed setting with 100 devices and 10 devices selected at random each time.

has about 20mm higher PA-MPJPE and MPJPE in the FastMETRO configuration. In contrast, in the TORE configuration, the gap is over 30mm higher PA-MPJPE and MPJPE. Compared to the naïve single-device training setting, experiment results have confirmed the superiority of larger corpus of training data with full Human3.6M dataset.

G Discussion on Multiple Local Epochs

# Local Rounds	Model	MPJPE	PA-MPJPE	Raw Data Local?
1	FM-S	59.31	41.48	✓
2	FM-S	59.28	42.42	✓
3	FM-S	60.49	42.43	✓
5	FM-S	59.52	42.07	✓
10	FM-S	60.03	42.96	✓
1	TORE-S	62.92	43.11	✓
2	TORE-S	61.84	43.27	✓
3	TORE-S	63.53	42.74	✓
5	TORE-S	63.41	43.18	✓
10	TORE-S	63.88	43.39	✓

Table 10: Effect of Multi Local Epochs Ablations on training HMR model with multiple local epochs in federated setting.

We also conduct ablation for model performance with multiple local epoches. The experiment is presented with `100/10` IID setting for H36M dataset, i.e. partitioned data with 100 client and sampled 10 at each round.

As shown in Tab 10, we found that, with more local epochs, the local model tends to overfit and result in less generalization capabilities and worse performance.

H Discussion on Data Distribution

Heterogeneous data distributions are common in real-world privacy-constrained model training since each user's local data distribution may be different due to user identity, environment, and user habits. Therefore, it is vital that we evaluate the model's performance under a heterogeneous data setting. In our experiments,

we measured the performance for heterogeneous data distribution of PCHMR. Here, we provide an additional discussion for each of the data distributions we discussed in the previous experiment section.

H.1 LDA Data Partition

One of the most utilized methodologies in FL to generate non-IID data is Latent Dirichlet Allocation (LDA) Hsu et al. (2019a). Specifically, LDA works by drawing the distribution $\boldsymbol{q}$ over N category class of labels from a Dirichlet distribution, i.e. $\boldsymbol{q} \sim \text{Dir}(\alpha \boldsymbol{p})$ where α is the concentration parameter controlling the skewness of the distribution and $\boldsymbol{p}$ is the vector characterizing prior distribution over N classes. This creates skewed datasets with approximation to real-world non-IID distribution.

However, in our case, the label category for the training set is hard to determine. In particular, each human mesh is, in fact, a unique label by definition, i.e. they all represent unique annotations of their character. Hence, two alternative methods for considering pseudo-labeling are labeling by subjects and scenes or labeling by the dataset composition. As many datasets only have one scene per object (in particular, pseudo-3D labeled datasets), labeling by scene and subjects seems infeasible. Therefore, we opted to label it with dataset composition instead and consider the category class as the sub-dataset from which the image originated.

On an IID dataset, applying LDA with higher α resulted in a similar distribution as uniform sampling, whereas applying LDA with lower α resulted in a much skewer non-IID distribution. Since our dataset is non-IID by dataset composition from the start, applying LDA partitioning on the data only increases the skewness of the data.

In our training dataset, as illustrated in Fig 8, applying LDA with low α creates a different distribution that is much more skewer, whereas applying LDA with high α retains the distribution that is similar to the original dataset composition.

In addition, as shown in Fig 9, LDA also creates the client dataset size distribution following the Dirichlet distribution, i.e. lower α creates a much skewer long-tail distribution in the number of samples assigned to each client. Essentially, LDA creates a more 'natural' distribution with individual client size variation, matching the real-world situation.

H.2 CNN Activation Clustering

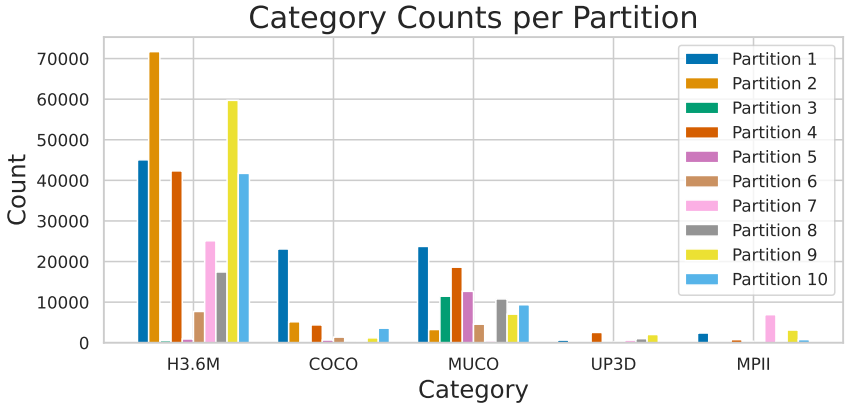
Another approach for doing pseudo-partition is considering natural clustering from the features. We used the same CNN feature extractor (ResNet-50 He et al. (2015)) from the PCHMR model to extract the feature from the last layer activation. The feature x_{out} extracted from the previous layer activation is then clustered to 100 pseudo partitions using MiniBatchedKMeans Sculley (2010). An example of the effects of clustering is illustrated in Fig. 10.

Upon inspection, we find that the feature extractor clusters also mimic the natural dataset composition relationship of the underlying dataset: the large groups of the clusters from the left consist mostly of the Human3.6M dataset, whereas the large cluster on the right consists of the pseudo-3D labeled dataset within the training dataset mixture. This demonstrate the effectiveness of this approach to generate a different data distribution for the given dataset.

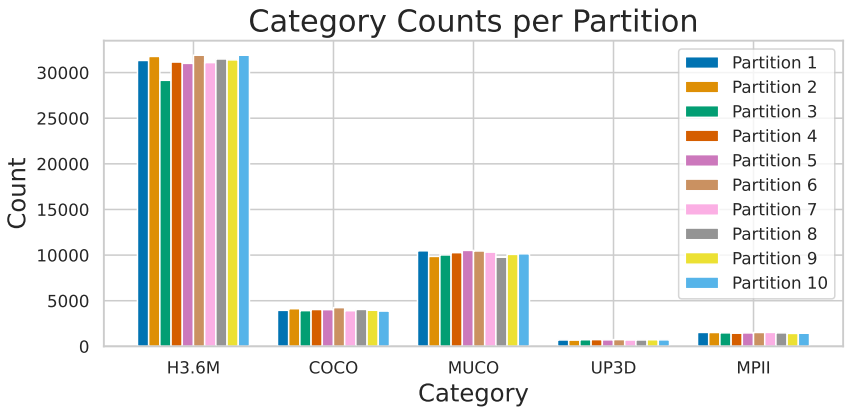
Under this setting, as shown in Tab. 11, our TORE Dou et al. (2023) model achieves the same level of accuracy compared to the centralized model, with a lower MPJPE error and slightly higher (+0.26mm) PA-MPJPE error. The FastMETRO Cho et al. (2022) variant also performs on par with the centralized model, achieving a lower MPJPE and slightly higher PA-MPJPE (+0.31mm).

H.3 Natural Partitioned In-the-wild Dataset

In addition, the centralized HMR model is usually trained on synthetic and in-door datasets due to the ease of gathering, not on the sparsely collected data from in-the-wild applications. These data usually exhibit a much different data distribution. We aim to further evaluate PCHMR under this real-world dataset with naturally



(a) LDA Data Distribution for $\alpha = 0.5$



(b) LDA Data Distribution for $\alpha = 1000$

Figure 8: Data Distribution of Applying LDA. This diagram illustrates the distribution of labels from each dataset using LDA with 5 clients under (a) $\alpha = 0.5$ and (b) $\alpha = 1000$. Note that the high α LDA retained a similar label distribution as the original category distribution, whereas the lower alpha introduced a much different category distribution.

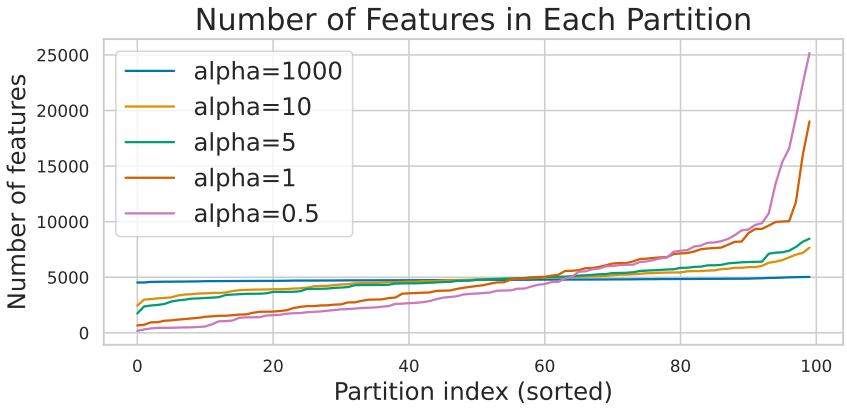
partitioned heterogeneous distribution labels. We utilize 3DPW von Marcard et al. (2018) as an example of such real-world collected in-the-wild data that presents much more challenging settings and is naturally partitioned with scene labels, which is demonstrated in Fig. 11. This makes them a good representation of the natural heterogeneous distribution that PCHMR deployment is likely to face in real-world applications.

H.4 Hetergenous Distribution Results

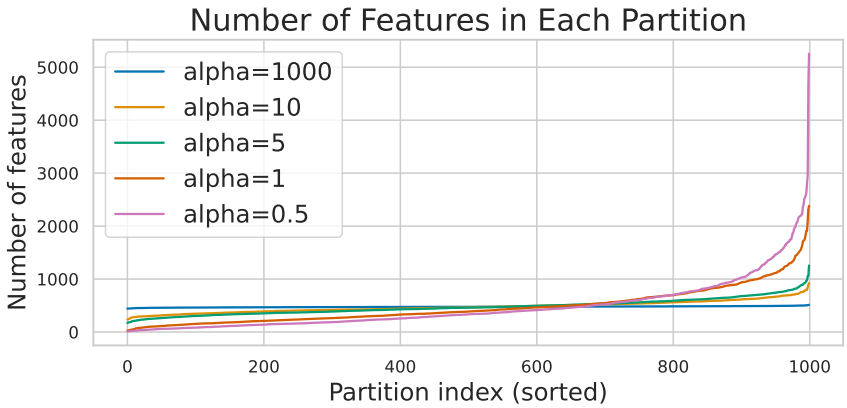
We then presents the complete results of heterogenous experiment at Tab 11 for reference.

I Performance of 3DPW Fine-tuning

One potential use-case of the PCHMR framework is performing privacy-constrained on-device fine-tuning to align the model with real-world data. We demonstrate this by pre-training our model with a series of datasets Ionescu et al. (2013); Lassner et al. (2017); Lin et al. (2014); Andriluka et al. (2014); Mehta et al. (2018) and using PCHMR to fine-tune on 3DPW Von Marcard et al. (2018) with 100 clients and 10 clients randomly sampled per round. As shown in Tab. 12, both fine-tuned options achieve significant performance gains



(a) Comparison of Data Assignment in LDA with 100 Partition



(b) Comparison of Data Assignment in LDA with 1000 Partition

Figure 9: LDA Partitioned Dataset Size for Individual Client. This figure illustrates the size of the dataset assigned to each client, sorted with the size of data for (a) 100 partition and (b) 1000 partition, using LDA. Note that the lower α creates a much skewer long tail distribution following the Dirichlet distribution that it is sampled from.

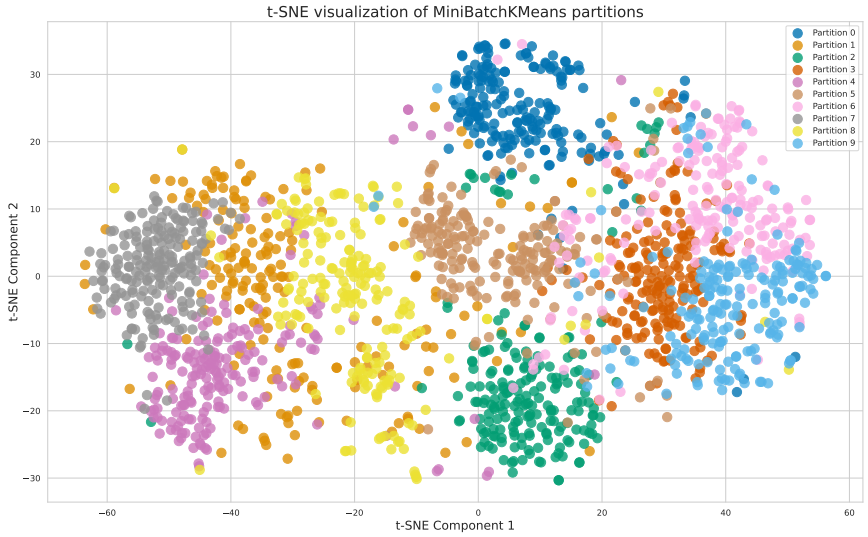


Figure 10: CNN Partitioned Data Visualisation. This diagram shows the selected datasets partitioned into 10 clients with KMeans and illustrated using the T-SNE van der Maaten & Hinton (2008). Each partition is sampled with 100 data points for ease of computation. The data points with the same color came from the same cluster.

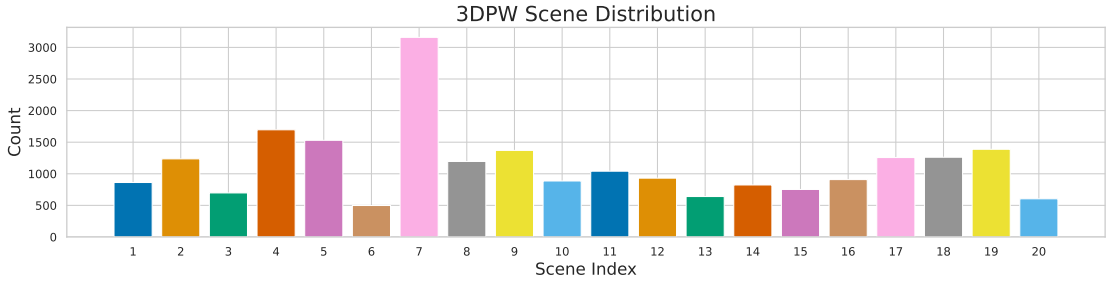


Figure 11: 3DPW Dataset Distribution. This diagram shows the natural distribution of different scenes in the 3DPW Von Marcard et al. (2018) dataset. Note that this is only the train set of the 3DPW dataset following the established data train-val partition.

compared to the non-fine-tuned model, verifying the necessity of fine-tuning. Fig. 12 also gives qualitative evidence for the on-device fine-tuning, where the reconstructed mesh accuracy visibly improves. Compared to centralized fine-tuning, the privacy-constrained fine-tuning results are comparable and in some metrics numerically slightly lower, but the differences are small. In summary, the on-device fine-tuning scheme keeps raw data local while offering competitive human reconstruction accuracy.

J On-Device Performance

We further assess the feasibility of the pipeline by measuring the performance of our PCHMR on a real edge device. We selected one of the representative embedded devices, Nvidia AGX Orin NVIDIA Corporation (2021), to run 5 rounds as one client in the 100- and 1000-client settings, following the experiment setting as outlined in the previous sections. The AGX Orin measurement provides a proxy for the runtime scale of edge deployment, but it is not a replacement for a full real-world multi-client deployment. With better capabilities hardware introduced on applications such as VR and AR, we expect the runtime profile to improve on newer hardware in the coming years.

The runtime and throughput results are shown in Tab 13. In the 1000 partition configuration, the edge device would need to train for 85 hours under the 50W power configuration. Nonetheless, the communication

Settings	Model	MPJPE	PA-MPJPE	Raw Data Local?
Centralized	FM-S	56.85	40.52	✗
$\alpha = 0.1$	FM-S	51.11	37.24	✓
$\alpha = 0.5$	FM-S	55.80	39.16	✓
$\alpha = 1$	FM-S	59.63	39.80	✓
$\alpha = 10$	FM-S	59.97	40.84	✓
$\alpha = 100$	FM-S	57.53	40.69	✓
$\alpha = 1000$	FM-S	59.80	41.80	✓
CNN Cluster	FM-S	56.84	40.83	✓
Centralized	TORE-S	63.88	41.99	✗
$\alpha = 0.1$	TORE-S	57.02	39.10	✓
$\alpha = 0.5$	TORE-S	58.91	40.94	✓
$\alpha = 1$	TORE-S	62.39	41.76	✓
$\alpha = 10$	TORE-S	60.96	42.25	✓
$\alpha = 100$	TORE-S	63.08	42.59	✓
$\alpha = 1000$	TORE-S	63.74	42.50	✓
CNN Cluster	TORE-S	63.14	42.25	✓

Table 11: Data Distribution Evaluation. This table shows the performance metrics (MPJPE, PAMPJPE) for different data distribution settings (with different LDA α and CNN Clustering label), evaluated with the Human3.6M Ionescu et al. (2013) dataset. The PCHMR's reconstruction accuracies are competitive and, in some simulated skewed partitions, lower-error than the centralized baseline; the mechanism behind these gains is not isolated here, supporting the robustness of our PCHMR approach under non-IID data distributions. All models (except centralized) in this table are trained in a distributed setting with 100 devices, with 10 devices selected at random each round.

Setting	Model	MPJPE	PA-MPJPE	MPVPE	Raw Data Local?
No Finetune	FM-S	116.35	70.20	132.96	✗
Centralized Finetune	FM-S	84.92	54.69	97.6	✗
PF Finetune	FM-S	84.66	54.78	97.42	✓
No Finetune	TORE-S	119.64	68.35	138.93	✗
Centralized Finetune	TORE-S	87.97	55.35	101.88	✗
PF Finetune	TORE-S	87.55	54.08	101.86	✓

Table 12: Privacy-Constrained Fine-Tuning Table. Results of privacy-constrained fine-tuning of models Dou et al. (2023); Cho et al. (2022) pre-trained on Ionescu et al. (2013); Lassner et al. (2017); Lin et al. (2014); Andriluka et al. (2014); Mehta et al. (2018) using PCHMR on the 3DPW dataset Von Marcard et al. (2018), compared with centralized fine-tuning with the same datasets and the model without fine-tuning. The privacy-constrained fine-tuning is conducted with 100 clients and 10 randomly sampled per round.

cost compared with both figures is relatively insignificant (42 seconds per round), assuming a 37.4 Mb/s broadband upload speed which follows the median uplink speed in the US by Ookla (2024) between June to Aug 2024. This shows that our PCHMR framework is within the runtime range of practical cross-device FL systems, given that normally the round completion times for practical FL can reach hundreds of hours Huba et al. (2022); Bonawitz et al. (2019).

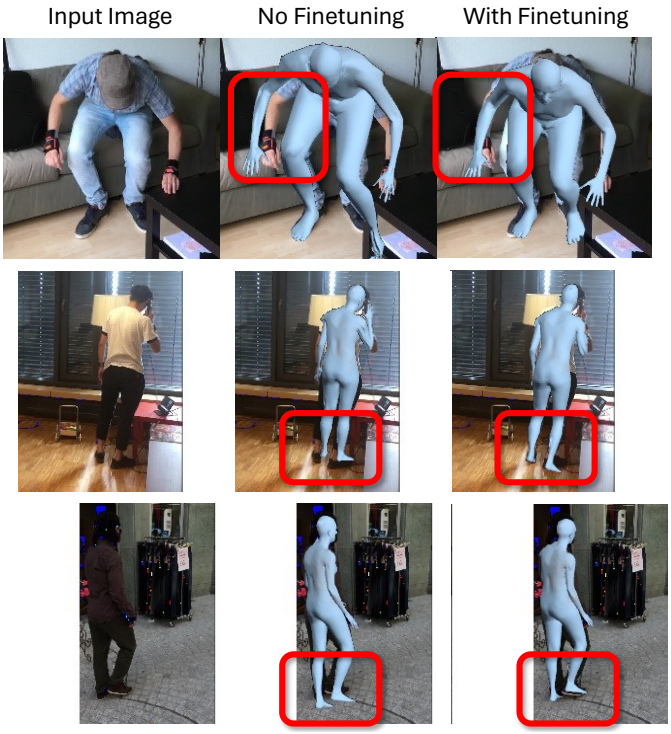


Figure 12: Qualitative results of Cho et al. (2022) before and after privacy-constrained fine-tuning on the 3DPW Von Marcard et al. (2018) dataset using the PCHMR framework.

Model	Device	100 Partition		1000 Partition	
		Throughput	Time	Throughput	Time
FM-S	Jetson AGX - 15W	2.46	1629.33	2.42	165.33
FM-S	Jetson AGX - 50W	8.01	499.67	7.81	51.23
FM-S	RTX A6000 Ada	56.59	70.7	48.0	8.33
TORE-S	Jetson AGX - 15W	3.43	1166.06	3.35	119.29
TORE-S	Jetson AGX - 50W	8.65	461.5	8.42	47.49
TORE-S	RTX A6000 Ada	50.87	78.67	44.40	9.02

Table 13: On Device Performance Table. This table shows the image throughput and total time used for training with 100 (4000 images) and 1000 (400 images) partitions of client data on different power profiles for Nvidia AGX Orion. We also compare the image throughput and total time with a desktop-level graphics card (RTX A6000 Ada). The unit for Throughput is images per second (img/s), and the unit for Time is seconds (s).