arXiv:2309.17448v3 [cs.CV] 28 Jul 2024

SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation

Zhongang Cai[*,1,2,3], **Wanqi Yin**[*,2,4], **Ailing Zeng**[5], **Chen Wei**[2], **Qingping Sun**[2], **Yanjun Wang**[2], **Hui En Pang**[1,2], **Haiyi Mei**[2], **Mingyuan Zhang**[1], **Lei Zhang**[5], **Chen Change Loy**[1], **Lei Yang**[†,2,3], **Ziwei Liu**[†,1]
[1] S-Lab, Nanyang Technological University, [2] SenseTime Research, [3] Shanghai AI Laboratory, [4] The University of Tokyo, [5] International Digital Economy Academy (IDEA)

Abstract

Expressive human pose and shape estimation (EHPS) unifies body, hands, and face motion capture with numerous applications. Despite encouraging progress, current state-of-the-art methods still depend largely on a confined set of training datasets. In this work, we investigate scaling up EHPS towards the first *generalist* foundation model (dubbed **SMPLer-X**), with up to ViT-Huge as the backbone and training with up to 4.5M instances from diverse data sources. With big data and the large model, SMPLer-X exhibits strong performance across diverse test benchmarks and excellent transferability to even unseen environments. *1) For the data scaling*, we perform a systematic investigation on 32 EHPS datasets, including a wide range of scenarios that a model trained on any single dataset cannot handle. More importantly, capitalizing on insights obtained from the extensive benchmarking process, we optimize our training scheme and select datasets that lead to a significant leap in EHPS capabilities. *2) For the model scaling,* we take advantage of vision transformers to study the scaling law of model sizes in EHPS. Moreover, our finetuning strategy turn SMPLer-X into *specialist* models, allowing them to achieve further performance boosts. Notably, our foundation model SMPLer-X consistently delivers state-of-the-art results on seven benchmarks such as AGORA (107.2 *mm* NMVE), UBody (57.4 *mm* PVE), EgoBody (63.6 *mm* PVE), and EHF (62.3 *mm* PVE without finetuning). [2]

1 Introduction

The recent progress in expressive human pose and shape estimation (EHPS) from monocular images or videos offers transformative applications for the animation, gaming, and fashion industries. This task typically employs parametric human models (*e.g.*, SMPL-X [51]) to adeptly represent the highly complicated human body, face, and hands. In recent years, a large number of diverse datasets have entered the field [5, 8, 63, 68, 39, 3, 14, 16, 64, 9], providing the community new opportunities to study various aspects such as capture environment, pose distribution, body visibility, and camera views. Yet, the state-of-the-art methods remain tethered to a limited selection of these datasets, creating a bottleneck in performance across varied scenarios and hindering the ability to generalize to unseen situations.

Our mission in this study is to explore existing data resources comprehensively, providing key insights crucial for establishing robust, universally applicable models for EHPS. Accordingly, we establish the first systematic benchmark for EHPS, utilizing 32 datasets and evaluating their performance

[*]Equal contributions. [†]Co-corresponding authors.
[2]Homepage: `https://caizhongang.github.io/projects/SMPLer-X/`.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

across five major benchmarks. We find that there are significant inconsistencies among benchmarks, revealing the overall complicated landscape of EHPS, and calling for data scaling to combat the domain gaps between scenarios. This detailed examination emphasizes the need to reassess the utilization of available datasets for EHPS, advocating for a shift towards more competitive alternatives that offer superior generalization capabilities, and highlights the importance of harnessing a large number of datasets to capitalize on their complementary nature.

Moreover, we systematically investigate the contributing factors that determine the transferability of these datasets. Our investigation yields useful tips for future dataset collection: 1) the more is not necessarily, the merrier: datasets do not have to be very large to be useful as long as they exceed approximately 100K instances based on our observation. 2) Varying indoor scenes is a good alternative if an in-the-wild (including outdoor) collection is not viable. 3) synthetic datasets, despite having traceable domain gaps, are becoming increasingly potent to a surprising extent. 4) Pseudo-SMPL-X labels are useful when ground truth SMPL-X annotations are unavailable.

Equipped with the knowledge procured from the benchmark, we exhibit the strength of massive data with SMPLer-X, a *generalist* foundation model that is trained using a diverse range of datasets and achieves exceptionally balanced results across various scenarios. To decouple from algorithmic research works, we design SMPLer-X with a minimalist mindset: SMPLer-X has a very simple architecture with only the most essential components for EHPS. We hope SMPLer-X could facilitate massive data and parameter scaling and serve as a baseline for future explorations in the field instead of a stringent investigation into the algorithmic aspect. Experiments with various data combinations and model sizes lead us to a well-rounded model that excels across all benchmarks that contests the community norm of limited-dataset training. Specifically, our foundation models demonstrate significant performance boost through both data scaling and model size scaling, reducing the mean primary errors on five major benchmarks (AGORA [50], UBody [39], EgoBody [68], 3DPW [58], and EHF [51]) from over 110 mm to below 70 mm (demonstrated in Fig. 1), and showcases impressive generalization capabilities by effectively transferring to new scenarios, such as DNA-Rendering [9] and ARCTIC [14].

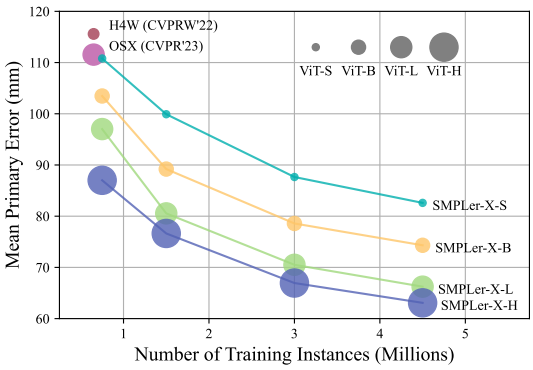

Figure 1: **Scaling up EHPS.** Both data and model scaling are effective in reducing mean errors on primary metrics across key benchmarks: AGORA [50], UBody [39], EgoBody [68], 3DPW [58] and EHF [51]. OSX [39] and H4W [46] are SOTA methods. Area of the circle indicates model size, with ViT variants as the reference (top right).

Furthermore, we validate the efficacy of finetuning our *generalist* foundation models to evolve into domain-specific *specialists*, delivering outstanding performance on all benchmarks. Specifically, we follow the same data selection strategy that empowers our specialist models to set new records on the AGORA leaderboard by being the first model to hit 107.2mm in NMVE (an 11.0% improvement) and achieving SOTA performance on EgoBody, UBody, and EHF.

Our contributions are three-fold. **1)** We build the first systematic and comprehensive benchmark on EHPS datasets, which provides critical guidance for scaling up the training data toward robust and transferable EHPS. **2)** We explore both data and model scaling in building the *generalist* foundation model that delivers balanced results across various scenarios and extends successfully to unseen datasets. **3)** We extend the data selection strategy to finetune the foundation model into potent *specialists*, catering to various benchmark scenarios.

2 Related Work

Expressive Human Pose and Shape Estimation (EHPS). Due to the erupting 3D virtual human research applications [66, 67, 21, 20, 7] and the parametric models (e.g., SMPL [42] and SMPL-X [51]), capturing the human pose and shape (HPS) [28, 33, 30, 31, 38, 59, 60], and additionally hands and face (EHPS) [51, 61, 10, 53, 70, 15, 56, 65] from images and videos have attracted increasing

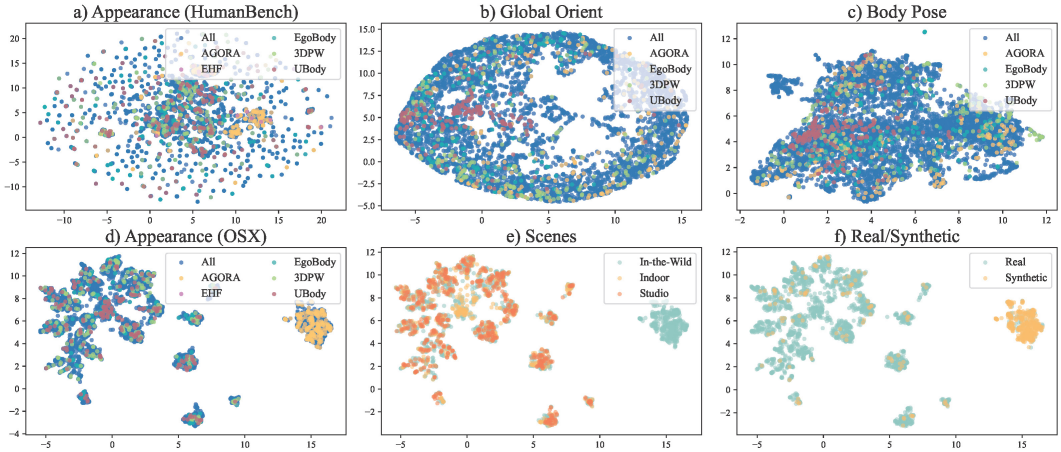

Figure 2: **Dataset attribute distributions.** a) and d) are image feature extracted by HumanBench [57] and OSX [39] pretrained ViT-L backbone. b) Global orientation (represented by rotation matrix) distribution. c) Body pose (represented by 3D skeleton joints) distribution. Both e) scenes and f) Real/Synthetic are drawn on the same distribution as d). All: all datasets. UMAP [43] dimension reduction is used with the x and y-axis as the dimensions of the embedded space (no unit).

attention. Optimization-based methods (e.g., SMPLify-X [51]) detect 2D features corresponding to the whole body and fit the SMPL-X model. However, they suffer from slow speed and are ultimately limited by the quality of the 2D keypoint detectors. Hence, learning-based models are proposed. One of the key challenges of EHPS is the low resolution of hands and face compared with the body-only estimation, making the articulated hand pose estimation and high-quality expression capture difficult. Accordingly, mainstream whole-body models first detect and crop the hands and face image patches, then resize them to higher resolutions and feed them into specific hand and face networks to estimate the corresponding parameters [10, 53, 70, 15, 56, 46, 65, 35]. Due to the highly complex multi-stage pipelines, they inevitably cause inconsistent and unnatural articulation of the mesh and implausible 3D wrist rotations, especially in occluded, truncated, and blurry scenes. Recently, OSX [39] proposes the first one-stage framework based on ViT-based backbone [13] to relieve the issues in previous multi-stage pipelines. This method provides a promising and concise way to scale up the model. However, they only use confined training datasets for a fair comparison and do not explore the combination of more data toward generalizable and precise EHPS.

Multi-dataset Training for Human-centric Vision. Recent efforts have been using multiple datasets in pretraining a general model for a wide range of downstream human-centric tasks. For example, HumanBench [57] leverages 37 datasets, whereas UniHCP [11] utilizes 33 datasets for tasks such as ReID, pedestrian detection, and 2D pose estimation. However, these works have only evaluated the efficacy of 2D tasks. Sárándi *et al.* [54] take advantage of 28 datasets in training a strong model for 3D keypoint detection, which recovers only the skeleton of subjects without estimating body shapes and meshes. Pang *et al.* [49] analyze 31 datasets for human pose and shape estimation (*i.e.*, SMPL estimation). However, hands and face estimation is not included, and only fewer than ten datasets are used concurrently in the most diverse training. This paper targets to scale training data and model size for EHPS, that simultaneously recovers the expressive pose and shape of the human body, hands, and face.

3 Benchmarking EHPS Datasets

3.1 Preliminaries

SMPL-X. We study expressive human pose and shape estimation via 3D parametric human model SMPL-X [51], which models the human body, hands, and face geometries with parameters. Specifically, our goal is to estimate pose parameters $\theta \in \mathbb{R}^{55\times 3}$ that include body, hands, eyes, and jaw poses; joint body, hands and face shape $\beta \in \mathbb{R}^{10}$, and facial expression $\psi \in \mathbb{R}^{10}$. The joint regressor $\mathcal{J}$ is used to obtain 3D keypoints from parameters via $R_\theta(\mathcal{J}(\beta))$ where R_θ is a transformation function along the kinematic tree.

Table 1: **Benchmarking EHPS datasets.** For each dataset, we train a model on its training set and evaluate its performance on the *val* set of AGORA and *testing* sets of UBody, EgoBody (EgoSet), 3DPW, and EHF. Datasets are then ranked by mean primary error (MPE). Top-1 values are bolded, and the rest of Top-5 are underlined. #Inst.: number of instances used in training. ITW: in-the-wild. EFT [27], NeuralAnnot (NeA) [47] and UP3D [34] produce pseudo labels.

						AGORA [50]		UBody [39]		EgoBody [68]		3DPW [58]		EHF [51]		
Dataset	#Inst.	Scene	Real/Synthetic	SMPL	SMPL-X	PVE↓	★	PVE↓	★	PVE↓	★	MPJPE↓	★	PVE↓	★	**MPE↓**
BEDLAM [5]	951.1K	ITW	Syn	-	Yes	164.7	4	132.5	8	109.1	2	**98.1**	1	**81.1**	1	**117.1**
SynBody [63]	633.5K	ITW	Syn	-	Yes	166.7	5	144.6	11	136.6	4	106.5	5	112.9	5	133.5
InstaVariety [29]	2184.8K	ITW	Real	NeA	-	195.0	9	125.4	4	140.1	9	100.6	3	110.8	4	134.3
GTA-Human II [8]	1802.2K	ITW	Syn	-	Yes	161.9	3	143.7	10	139.2	8	103.4	4	126.0	12	134.8
MSCOCO [40]	149.8K	ITW	Real	EFT	NeA	191.6	8	107.2	2	139.0	7	121.2	10	116.3	7	135.0
EgoBody-MVSet [68]	845.9K	Indoor	Real	Yes	Yes	190.9	7	191.4	18	127.0	3	99.2	2	101.8	2	142.1
AGORA [50]	106.7K	ITW	Syn	Yes	Yes	**124.8**	1	128.4	6	138.4	6	131.1	12	164.6	24	145.4
Egobody-EgoSet [68]	90.1K	Indoor	Real	Yes	Yes	207.1	15	126.8	5	**103.1**	1	134.4	18	121.4	10	147.5
RICH [22]	243.4K	ITW	Real	-	Yes	195.6	10	168.1	15	137.9	5	115.5	8	127.5	13	148.9
MPII [2]	28.9K	ITW	Real	EFT	NeA	202.1	11	123.9	3	155.5	15	131.9	14	140.8	16	150.8
MuCo-3DHP [45]	465.3K	ITW	Real	Yes	-	187.7	6	185.4	17	146.4	12	119.4	9	134.7	15	154.7
PROX [19]	88.5K	Indoor	Real	-	Yes	204.1	13	180.3	16	151.8	13	132.5	17	122.5	11	158.2
UBody [39]	683.3K	ITW	Real	-	Yes	207.0	14	**78.7**	1	145.6	11	149.4	23	132.1	14	158.5
SPEC [32]	72.0K	ITW	Syn	Yes	-	161.5	2	146.1	12	154.8	14	139.7	21	197.8	27	160.0
CrowdPose [36]	28.5K	ITW	Real	NeA	-	207.1	16	129.8	7	156.9	16	156.3	25	154.5	22	160.9
MPI-INF-3DHP [44]	939.8K	ITW	Real	NeA	NeA	221.5	20	166.7	14	142.7	10	131.6	13	155.5	23	163.6
HumanSC3D [17]	288.4K	Studio	Real	-	Yes	215.2	18	237.8	22	167.3	17	113.0	7	107.1	3	168.1
PoseTrack [1]	28.5K	ITW	Real	EFT	-	218.1	19	161.0	13	180.8	21	150.2	24	149.9	21	172.0
BEHAVE [3]	44.4K	Indoor	Real	Yes	-	208.3	17	205.8	20	175.8	19	132.0	15	145.0	18	173.4
CHI3D [16]	252.4K	Studio	Real	-	Yes	203.3	12	264.7	25	175.7	18	122.6	11	121.0	9	177.5
Human3.6M [23]	312.2K	Studio	Real	Yes	NeA	226.0	21	276.1	26	200.6	24	112.3	6	120.8	8	187.2
DNA-R-HiRes [9]	998.1K	Studio	Real	-	Yes	230.0	22	278.2	27	179.2	20	134.5	19	149.7	20	194.3
3DPW [58]	22.7K	ITW	Real	Yes	NeA	234.0	23	259.3	23	192.6	23	140.6	22	142.9	17	207.2
ARCTIC [14]	1539.1K	Studio	Real	-	Yes	308.5	29	200.7	19	186.4	22	202.5	26	182.5	25	216.1
DNA-R [9]	3992.0K	Studio	Real	-	Yes	274.7	26	341.5	30	214.4	27	138.4	20	115.5	6	216.9
UP3D [34]	7.1K	ITW	Real	UP3D	-	257.5	24	224.1	21	216.6	28	211.5	27	194.8	26	220.9
Talkshow [64]	3326.9K	Indoor	Real	-	Yes	286.4	27	133.2	9	203.6	25	291.3	29	201.9	28	223.3
FIT3D [18]	1779.3K	Studio	Real	-	Yes	329.7	30	404.0	31	213.8	26	132.1	16	148.1	19	245.5
MTP [48]	3.2K	ITW	Real	Yes	Yes	272.7	25	284.9	28	273.2	29	265.2	28	244.6	29	268.1
OCHuman [69]	2.5K	ITW	Real	EFT	-	307.1	28	263.3	24	279.3	30	293.4	30	281.7	30	285.0
LSPET [25]	2.9K	ITW	Real	EFT	-	365.7	31	292.6	29	340.1	31	339.8	31	316.3	31	330.9
SSP3D [55]	311	ITW	Real	Yes	-	549.8	32	522.4	32	548.1	32	439.0	32	539.5	32	519.8

Evaluation Metrics. We use standard metrics for EHPS. PVE (per-vertex error) and MPJPE (mean per-joint position error) measure the mean L2 error for vertices and regressed joints, respectively. The "PA" prefix indicates Procrutes Alignment is conducted before error computation. AGORA Leaderboard [50] introduces NMVE (normalized mean vertex error) and NMJE (normalized mean joint error) that take detection performance F1 score into consideration. Moreover, we propose MPE (mean primary error) that takes the mean of multiple primary metrics (MPJPE for 3DPW [58] test, and PVE for AGORA, UBody, EgoBody, and EHF) to gauge generalizability. All errors are reported in millimeters (mm).

3.2 Overview of Data Sources

In this work, we study three major types of datasets. 1) motion capture datasets that leverage optical [23, 16, 14, 17, 18, 44] or vision-based [68, 19, 9, 7] multi-view motion capture systems, are typically collected in a studio environment. However, it is possible to include an outdoor setup, or utilize additional sensors such as IMUs [58]. These datasets generally provide high-quality 3D annotations but are less flexible due to physical constraints, especially those built with immobile capture systems that require accurate sensor calibrations. 2) pseudo-annotated datasets [40, 1, 36, 39, 45, 29, 69, 2, 25, 48, 64, 55] that re-annotate existing image datasets with parametric human annotations [26, 47, 39]. These datasets take advantage of the diversity of 2D datasets, and the pseudo-3D annotations, albeit typically not as high-quality, have been proven effective [49, 33, 26]. 3) synthetic datasets [5, 8, 32, 50, 63] that are produced with renderings engines (*e.g.*, Unreal Engine). These datasets produce the most accurate 3D annotations and can easily scale up with high diversity. However, the synthetic-real gap is not fully addressed. Key attributes of the datasets are included in Table 1.

To evaluate the EHPS capability across diverse scenarios, we select multiple key datasets to form a comprehensive benchmark. They should possess the desirable traits such as 1) having accurate SMPL or SMPL-X annotations, 2) being representative of certain aspects of real-life scenarios, 3) being widely used, but this requirement is relaxed for the new datasets which are released within two years, and 4) has a clearly defined test set. To this end, five datasets (AGORA [50], UBody [39], EgoBody [68], 3DPW [58], and EHF [51]) representing different aspects are selected as the evaluation datasets. We briefly introduce these five datasets and the rest in the Supplementary Material. **AGORA**

is the most widely-used benchmark for SMPL-X evaluation. It is a synthetic dataset featuring diverse subject appearances, poses, and environments with high-quality annotation. We evaluate on both validation and test set (leaderboard) as the latter has a monthly limit of submissions. **UBody** is the latest large-scale dataset with pseudo-SMPL-X annotations that covers fifteen real-life scenarios, such as talk shows, video conferences, and vlogs, which primarily consist of the upper body in images. We follow the intra-scene protocol in training and testing, where all scenarios are seen. **EgoBody** captures human motions in social interactions in 3D scenes with pseudo-SMPL-X annotations. It comprises a first-person egocentric set (EgoSet) and a third-person multi-camera set (MVSet). We test on the EgoSet with heavy truncation and invisibility. **3DPW** is the most popular in-the-wild dataset with SMPL annotations. Since SMPL-X annotation is not available, we map SMPL-X keypoints and test on 14 LSP [24] keypoints following the conventional protocol [28, 33]. **EHF** is a classic dataset with 100 curated frames of one subject in an indoor studio setup, with diverse body poses and especially hand poses annotated in SMPL-X vertices. It has a test set but no training or validation sets. Hence, it is only used to evaluate cross-dataset performance.

Besides being popular or the latest evaluation sets for EHPS, we further analyze if these five datasets collectively provide wide coverage of existing datasets. In Fig. 3, we randomly downsample all datasets to equal length (1K examples) and employ UMAP [43] to visualize several key aspects. We use pretrained ViT-L from HumanBench [57] and OSX [39] to process patch tokens flattened as feature vectors from images cropped by bounding boxes. HumanBench is trained for various human-centric tasks (*e.g.*, Re-ID, part segmentation, and 2D pose estimation), whereas OSX is an expert model on EHPS. As for global orientation, it is closely associated with camera pose as we convert all data into the camera coordinate frame; we plot its distribution by using flattened rotation matrix representations. Moreover, we follow [52, 8, 49] to represent poses as 3D keypoints regressed from the parametric model. Specifically, we flatten 21 SMPL-X body keypoints, and 15 hand keypoints from each hand, regressed with zero parameters except for the body pose and hand poses. It is shown that 1) the five benchmark datasets have varied distribution, which is expected due to their different designated purposes, and 2) collectively, the five datasets provide a wide, near-complete coverage of the entire dataset pool.

3.3 Benchmarking on Individual Datasets

In this section, we aim to benchmark datasets and find those that do well in various scenarios. To gauge the performance of each dataset, we train a SMPLer-X model with the training set of that dataset and evaluate the model on the *val/testing* sets of five evaluation datasets: AGORA, UBody, EgoBody, 3DPW, and EHF. Here, the benchmarking model is standardized to use ViT-S as the backbone, trained on 4 V100 GPUs for 5 epochs with a total batch size of 128 and a learning rate of 1×10^{-5}. The dataset preprocessing details are included in the Supplementary Material.

In Table 1, we report the primary metrics (Sec. 3.1) and ranking of the 32 datasets. The complete results in the Supplementary Material. We also compute the mean primary error (MPE) to facilitate easy comparison between individual datasets. Note that for AGORA, UBody, EgoBody, and 3DPW, their performances on their own test set are excluded from computing MPE. This is because in-domain evaluation results are typically much better than cross-domain ones, leading to significant error drops. In addition, note that there are datasets designed for specific purposes (*e.g.*, Talkshow [64] for gesture generation, DNA-Rendering [9] for human NeRF reconstruction), being ranked lower on our benchmark, which focuses on EHPS (a perception task) does not reduce their unique values and contributions to the computer vision community.

From the benchmark, we observe models trained on a single dataset tend to perform well on the same domain but often cannot do well on other domains. For example, the model trained on AGORA is ranked 1st on AGORA (val), but 6th on UBody, 6th on EgoBody, 12th on 3DPW, and 24th on EHF. This observation indicates that 1) the test scenarios are diverse, showcasing the challenging landscape of EHPS, and 2) data scaling is essential for training a robust and transferable model for EHPS due to significant gaps between different domains.

3.4 Analyses on Dataset Attributes

In this section, we study attributes that contribute to generalizability. However, it is important to acknowledge that such analyses are not a straightforward task: the attributes often exhibit coupled

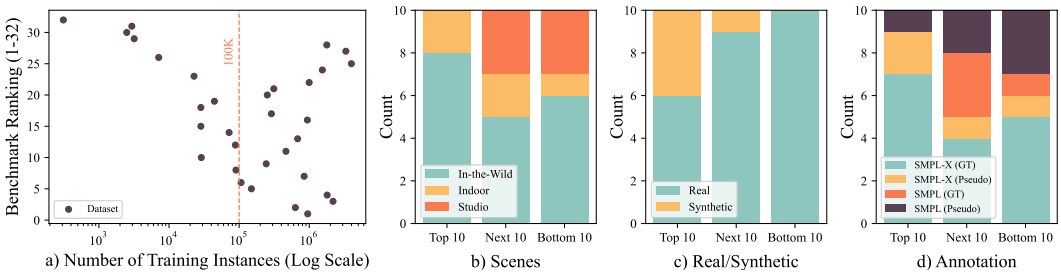

Figure 3: **Analysis on dataset attributes.** We study the impact of a) the number of training instances, b) scenes, c) real or synthetic appearance, and d) annotation type, on dataset ranking in Table 1.

effects. Consequently, counter-examples are inevitable (*e.g.*, we observe that InstaVariety, an in-the-wild dataset, demonstrates strong performance, whereas LSPET, another in-the-wild dataset, does not perform as well). Despite the challenges in pinpointing the exact factors that determine the success of an individual dataset, we adopt a collective perspective and aim to identify general trends with several key factors [49, 41, 8] in Fig. 3, and discussed below.

First, Fig. 3a) shows that the performance of a dataset (in terms of ranking) is not strongly associated with the number of training instances once the instance number exceeds approximately 100K. Although a very small amount of training data is insufficient to train a strong model, having an exceedingly large amount of data does not guarantee good performance either. For example, MSCOCO only comprises 149.8K training instances but achieves a higher ranking compared to datasets with 10$\times$ larger scales. This may be attributed to the diverse appearance and complex scenes present in the MSCOCO dataset. Hence, it would be more cost-effective to channel resources to improve diversity and quality, when the dataset has become adequately large.

Second, we categorize datasets into 1) in-the-wild, which contains data from diverse environments; 2) indoor with several scenes; 3) studio, which has a fixed multi-view setup. Particularly, Fig. 3b) shows that the top 10 are mostly in-the-wild datasets, indoor datasets concentrate in the top 20 and the studio dataset tends to be ranked lower in the benchmark. Moreover, Fig. 2e) illustrates that in-the-wild datasets exhibit the most diverse distribution, covering both indoor and studio datasets. Indoor datasets display a reasonable spread, and studio datasets have the least diversity. Our findings validate previous studies that suggest an indoor-outdoor domain gap [27]. Differing from Pang *et al.* [49], which does not differentiate between indoor and studio datasets, we argue that categorizing all datasets collected indoors into a single class oversimplifies the analysis. For example, consider EgoBody [68] and Human3.6M [23]. Both datasets does not have outdoor data; however, EgoBody consists of a wide variety of indoor scenes, whereas Human3.6M consists of only one scene, which may contribute to the better ranking of EgoBody compared to Human3.6M. Hence, this suggests that in-the-wild data collection is the most ideal, but diversifying indoor scenes is the best alternative.

Third, most of the five contemporary synthetic datasets [5, 63, 50, 32, 8] demonstrate surprising strength and are ranked highly in Fig. 3c). It is worth noting that four (UBody, EgoBody, 3DPW, and EHF) of the five evaluation benchmarks used are real datasets, indicating that knowledge learned from synthetic data is transferable to real scenarios. To explain this observation, we take a close look at Fig. 2f): although real and synthetic datasets do not have extensive overlap, synthetic data possesses two ideal characteristics. First, there is a high overlap between real and synthetic data at the rightmost cluster. Referring to Fig. 2e), which is drawn from the same distribution, we find that this cluster primarily represents in-the-wild data. Therefore, synthetic data includes a substantial number of in-the-wild images that closely resemble real in-the-wild scenarios. Second, synthetic data also have scatters of image features on other clusters, indicating that synthetic data provides coverage to some extent for various real-world scenarios.

Fourth, Fig. 3d) reveals that a dataset can be valuable with accurate or pseudo-SMPL-X annotations, as they constitute the most of the top 10 datasets. A prominent example is InstaVariety [29], which has only pseudo-SMPL-X annotation produced by NeuralAnnot [47], yet, is ranked third in our benchmark. However, due to the differences in parameter spaces, SMPL annotations are less effective: it is observed that datasets with SMPL annotations tend to cluster in the lower bracket of the benchmark, especially those with pseudo-SMPL annotations. This observation suggests that SMPL-X

Table 2: **Foundation Models.** We study the scaling law of the amount of data and the model sizes. The metrics are MPJPE for 3DPW, and PVE for other evaluation benchmarks. Foundation models are named "SMPLer-X-MN", where M indicates the size of ViT backbone (S, B, L, H), N is the number of datasets used in the training. FPS: inference speed (frames per second) on a V100 GPU. MPE: mean primary error. AGORA uses the validation set, and EgoBody uses the EgoSet.

#Datasets	#Inst.	Model	#Param.	FPS	AGORA [50]	EgoBody [68]	UBody [39]	3DPW [58]	EHF [51]	MPE
5	0.75M	SMPLer-X-S5	32M	36.2	119.0	114.2	110.1	110.2	100.5	110.8
10	1.5M	SMPLer-X-S10	32M	36.2	116.0	88.6	107.7	97.4	89.9	99.9
20	3.0M	SMPLer-X-S20	32M	36.2	109.2	84.3	70.7	87.5	86.6	87.7
32	4.5M	SMPLer-X-S32	32M	36.2	105.2	82.5	68.1	83.2	74.1	82.6
5	0.75M	SMPLer-X-B5	103M	33.1	102.7	108.1	105.8	104.8	96.1	103.5
10	1.5M	SMPLer-X-B10	103M	33.1	97.8	76.4	107.3	89.9	74.7	89.2
20	3.0M	SMPLer-X-B20	103M	33.1	95.6	75.5	65.3	83.5	73.0	78.6
32	4.5M	SMPLer-X-B32	103M	33.1	88.0	72.7	63.3	80.3	67.3	74.3
5	0.75M	SMPLer-X-L5	327M	24.4	88.3	98.7	110.8	97.8	89.5	97.0
10	1.5M	SMPLer-X-L10	327M	24.4	82.6	69.7	104.0	82.5	64.0	80.6
20	3.0M	SMPLer-X-L20	327M	24.4	80.7	66.6	61.5	78.3	65.4	70.5
32	4.5M	SMPLer-X-L32	327M	24.4	74.2	62.2	57.3	75.2	62.4	66.2
5	0.75M	SMPLer-X-H5	662M	17.5	89.0	87.4	102.1	88.3	68.3	87.0
10	1.5M	SMPLer-X-H10	662M	17.5	81.4	65.7	100.7	78.7	**56.6**	76.6
20	3.0M	SMPLer-X-H20	662M	17.5	77.5	63.5	59.9	**74.4**	59.4	67.0
32	4.5M	SMPLer-X-H32	662M	17.5	**69.5**	**59.5**	**54.5**	75.0	56.8	**63.1**

annotations are critical to EHPS; fitting pseudo labels is a useful strategy even if they could be noisy. Moreover, using SMPL labels effectively for SMPL-X estimation remains a challenge.

4 Scaling up EHPS

4.1 Model Architectures

Catering to our investigation, we design a minimalistic framework (dubbed SMPLer-X) that only retains the most essential parts for two reasons. First, it must be scalable and efficient as we train with a large amount of data. Second, we aim to create a framework that is decoupled from specific algorithm designs, providing a clean foundation for future research. To this end, SMPLer-X consists of three parts: a *backbone* extracts image features, which we employ Vision Transformer [13] for its scalability; a *neck* that predicts bounding boxes and crop regions of interest from the feature map for hands and face; regression *heads* that estimate parameters for each part. Note that SMPLer-X does not require third-party detectors [53], cross-part feature interaction modules [10, 15], projection of coarse SMPL-X estimations [65], or a heavy decoder [39]. As the design of SMPLer-X is not the focus of our investigation, more details are included in the Supplementary Material.

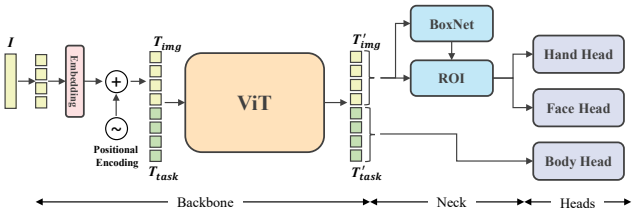

Figure 4: **Architecture of SMPLer-X**, which upholds the idea that "simplicity is beauty". SMPLer-X contains a backbone that allows for easy investigation on model scaling, a neck for hand and face feature cropping, and heads for different body parts. Note that we wish to show in this work that model and data scaling are effective, even with a straightforward architecture.

4.2 Training the Generalist Foundation Models

The SOTA methods [39, 46] usually train with only a few (*e.g.*, MSCOCO, MPII, and Human3.6M) datasets, whereas we investigate training with many more datasets. However, we highlight that the dataset benchmark in Table 1 cannot be used: selecting datasets based on their performance on the test sets of the evaluation benchmarks leaks information about the test sets. Hence, we construct another dataset benchmark in the Supplementary Material, that ranks individual datasets on the *training* set of the major EHPS benchmarks. We use four data amounts: 5, 10, 20, and 32 datasets as the training set, with a total length of 0.75M, 1.5M, 3.0M, and 4.5M instances. We always prioritize higher-ranked

Table 3: **AGORA test set.** † denotes the methods that are finetuned on the AGORA training set. ∗denotes the methods that are trained on AGORA training set only.

	NMVE↓ (*mm*)		NMJE↓ (*mm*)		MVE↓ (*mm*)					MPJPE↓ (*mm*)				
Method	All	Body	All	Body	All	Body	Face	LHand	RHand	All	Body	Face	LHand	RHhand
BEDLAM [5]	179.5	132.2	177.5	131.4	131.0	96.5	25.8	38.8	39.0	129.6	95.9	27.8	36.6	36.7
Hand4Whole [46]†	144.1	96.0	141.1	92.7	135.5	90.2	41.6	46.3	48.1	132.6	87.1	46.1	44.3	46.2
BEDLAM [5]†	142.2	102.1	141.0	101.8	103.8	74.5	**23.1**	**31.7**	**33.2**	102.9	74.3	**24.7**	**29.9**	**31.3**
PyMaF-X [65]†	141.2	94.4	140.0	93.5	125.7	84.0	35.0	44.6	45.6	124.6	83.2	37.9	42.5	43.7
OSX [39] ∗	130.6	85.3	127.6	83.3	122.8	80.2	36.2	45.4	46.1	119.9	78.3	37.9	43.0	43.9
HybrIK-X [35]	120.5	73.7	115.7	72.3	112.1	68.5	37.0	46.7	47.0	107.6	67.2	38.5	41.2	41.4
SMPLer-X-L20	133.1	88.1	128.9	84.6	123.8	81.9	37.4	43.6	44.8	119.9	78.7	39.5	41.4	44.8
SMPLer-X-L32	122.8	80.3	119.1	77.6	114.2	74.7	35.1	41.3	42.2	110.8	72.2	36.7	39.1	40.1
SMPLer-X-L20†	**107.2**	**68.3**	**104.1**	**66.3**	**99.7**	**63.5**	29.9	39.1	39.5	**96.8**	**61.7**	31.4	36.7	37.2

Table 4: **AGORA Val set**. † and ∗ are finetuned on the AGORA training set, and trained on the AGORA training set only, respectively.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]†	73.2	9.7	4.7	183.9	72.8	81.6
OSX [39]	69.4	11.5	4.8	168.6	70.6	77.2
OSX [39]∗	45.0	**8.5**	3.9	79.6	48.2	37.9
SMPLer-X-B1∗	48.9	8.6	4.0	86.1	51.5	41.2
SMPLer-X-L20	48.6	8.9	4.0	80.7	51.0	41.3
SMPLer-X-L32	45.1	8.7	**3.8**	74.2	47.8	38.7
SMPLer-X-L20†	**39.1**	9.3	**3.8**	**62.5**	**42.3**	**32.8**

Table 5: **EHF**. As EHF does not have a training set to benchmark datasets, we do not perform finetuning. Moreover, EHF is not seen in our training and can be used to validate our foundation models' transferability.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	50.3	**10.8**	5.8	76.8	**39.8**	26.1
OSX [39]	48.7	15.9	6.0	70.8	53.7	26.4
SMPLer-X-L20	37.8	15.0	5.1	65.4	49.4	17.4
SMPLer-X-L32	**37.1**	14.1	**5.0**	**62.4**	47.1	**17.0**

datasets. To prevent larger datasets from shadowing smaller datasets, we adopt a balanced sampling strategy. Specifically, all selected datasets are uniformly upsampled or downsampled to the same length and add up to the designated total length. To facilitate training, we follow OSX [39] to use AGORA, UBody, MPII, 3DPW, Human3.6M in COCO-format [40], and standardize all other datasets into the HumanData [12] format. We also study four ViT backbones of different sizes (ViT-Small, Base, Large and Huge), pretrained by ViTPose [62]. The training is conducted on 16 V100 GPUs, with a total batch size of 512 (256 for ViT-Huge) for 10 epochs. More training details such as adapting SMPL or gendered SMPL-X in the training are included in the Supplementary Material.

In Table 2, we show experimental results with a various number of datasets and foundation model sizes. Foundation models are named "SMPLer-X-MN", where M can be S, B, L, H that indicates the size of the ViT backbone, and N indicates the number of datasets used in the training. For example, SMPLer-X-L10 means the foundation model takes ViT-L as the backbone, and is trained with Top 10 datasets (ranked according to the individual dataset performance on the training sets of the key evaluation benchmarks). It is observed that **1)** more training data (data scaling) leads to better performance in terms of MPE. The model performance improves gradually as the number of training datasets increases. However, besides the increment in training instances, more datasets provide a richer collection of diverse scenarios, which we argue is also a key contributor to the performance gain across evaluation benchmarks. **2)** A larger foundation model (model scaling) performs better at any given amount of data. However, the marginal benefits of scaling up decrease beyond model size L. Specifically, ViT-H has more than twice the parameters than ViT-L, but the performance gain is not prominent. **3)** The foundation model always performs better than in-domain training on a single training set. For example, SMPLer-X-B20, performs better on the validation set of AGORA, and test sets of UBody, EgoBody, and 3DPW, than models trained specifically on the corresponding training set in Table 1. This is useful for real-life applications: instead of training a model for each of the user cases, a generalist foundation model contains rich knowledge to be a one-size-fits-all alternative.

Table 10: **3DPW.** ‡ denotes the methods that use a head for SMPL regression. † and ∗ are finetuned on the 3DPW training set and trained on 3DPW training set only, respectively. Unit: *mm*.

Method	MPJPE	PA-MPJPE
Body-only (SMPL) Methods		
OSX-SMPL [39]‡∗	74.7	45.1
HybrIK [37]	71.6	**41.8**
CLIFF [38]	**68.0**	43.0
Whole-Body (SMPL-X) Methods		
Hand4Whole [46]	86.6	54.4
ExPose [10]	93.4	60.7
OSX [39]†	86.2	60.6
SMPLer-X-B1∗	95.6	67.6
SMPLer-X-L20	78.3	52.1
SMPLer-X-L32	**75.2**	**50.5**
SMPLer-X-L20†	76.8	51.5

Table 6: **UBody.** † denotes the methods that are finetuned on the UBody training set. ∗ denotes the methods that are trained on UBody training set only.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
PIXIE [15]	61.7	12.2	4.2	168.4	55.6	45.2
Hand4Whole [46]	44.8	8.9	2.8	104.1	45.7	27.0
OSX [39]	42.4	10.8	2.4	92.4	47.7	24.9
OSX [39]†	42.2	**8.6**	**2.0**	81.9	41.5	**21.2**
SMPLer-X-B1∗	38.5	10.8	3.0	64.8	45.4	22.3
SMPLer-X-L20	33.2	10.6	2.8	61.5	43.3	23.1
SMPLer-X-L32	**30.9**	10.2	2.7	**57.3**	**39.2**	21.6
SMPLer-X-L-20†	31.9	10.3	2.8	57.4	40.2	21.6

Table 7: **EgoBody-EgoSet.** † denotes the methods that are finetuned on the EgoBody-EgoSet training set. ∗ denotes the methods that are trained on EgoBody-EgoSet training set only.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	58.8	**9.7**	3.7	121.9	50.0	42.5
OSX [39]	54.6	11.6	3.7	115.7	50.6	41.1
OSX [39]†	45.3	10.0	3.0	82.3	46.8	35.2
SMPLer-X-B1∗	56.1	10.7	3.5	87.2	49.4	34.9
SMPLer-X-L20	38.9	9.9	3.0	66.6	42.7	31.8
SMPLer-X-L32	**36.3**	9.8	**2.9**	**62.2**	**41.4**	**30.7**
SMPLer-X-L20†	37.8	9.9	**2.9**	63.6	42.5	30.8

Table 8: **ARCTIC.** † and ∗ denote the methods that are finetuned on the ARCTIC training set and trained on the ARCTIC training set only, respectively.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	63.4	18.1	4.0	136.8	54.8	59.2
OSX [39]	56.9	**17.5**	3.9	102.6	56.5	44.6
OSX [39]†	33.0	18.8	3.3	58.4	39.4	30.4
SMPLer-X-B1∗	45.2	18.9	3.4	66.6	42.5	34.0
SMPLer-X-L10	46.9	18.1	**2.3**	76.9	50.8	33.2
SMPLer-X-L32	**29.4**	18.9	2.7	**48.6**	**38.8**	**26.8**
SMPLer-X-L10†	33.1	19.0	2.7	54.9	40.1	27.3

Table 9: **DNA-Rendering-HiRes**. † and ∗ are finetuned on the DNA-Rendering-HiRes training set and trained on the DNA-Rendering-HiRes training set only, respectively.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	62.8	11.0	4.2	111.4	56.4	52.6
OSX [39]	59.9	10.6	4.3	105.7	55.0	52.5
OSX [39]†	43.5	7.5	3.5	67.1	43.3	38.2
SMPLer-X-B1∗	45.6	7.5	3.4	63.2	40.7	34.2
SMPLer-X-L20	44.4	11.1	4.5	77.7	47.5	43.2
SMPLer-X-L32	**35.8**	**7.2**	**3.2**	**54.4**	**36.7**	**34.0**
SMPLer-X-L20†	37.9	7.3	3.4	56.5	38.4	34.9

Besides the errors on key benchmarks, we also report the inference speed (in terms of FPS, or frames per second) of the SMPLer-X model family in Table 2. The testing is conducted on a single V100 GPU with batch size 1, excluding data loading. SMPLer-X family is faster than OSX (12.2 FPS on a single V100 GPU) using the same test setting, and the smaller versions such as SMPLer-X-S and SMPLer-X-B can achieve real-time performance, with SMPLer-X-L on the verge of achieving real-time speeds. The high inference speed is attributed to the minimalistic architecture of SMPLer-X, which only retains the most essential components for EHPS.

Moreover, we show detailed by-part results of body, hands, and face on main benchmarks such as AGORA test set (Table 3), AGORA validation set (Table 4), UBody (Table 6), EgoBody-EgoSet (Table 7) and EHF (Table 5). We also compare our results with whole-body methods on 3DPW (Table 10). We highlight that the foundation models show strong and balanced performances on all benchmarks.

Furthermore, we evaluate the transferability of our foundation models on two more benchmarks: ARCTIC (Table 8) and DNA-Rendering (Table 8). ARCTIC features complicated hand-object interaction with whole-body annotations, and DNA-Rendering includes diverse subjects, motions, and garments. Note that ARCTIC is not seen by foundation models trained on Top 10 datasets, and DNA-Rendering is not seen by foundation models trained on Top 20 datasets. The foundation models, however, achieve much better performance than SOTAs with conventional data sampling strategies.

In addition, we compare our foundation model with SOTA methods, such as Hand4Whole [46] and OSX [39] in various scenarios in Fig. 5. These scenarios feature challenging aspects such as heavy truncation (from only half of the body visible to only the arms visible), difficult body poses in diverse backgrounds, and rare camera angles (extremely high or low elevation angles). SMPLer-X demonstrates the strength of massive training data and consistently produces robust estimations.

4.3 Finetuning the Specialists

Training the foundation model with a large number of data is expensive. For example, SMPLer-X-H20 takes more than 400 GPU hours to train. Hence, it is critical to investigate finetuning strategies that allow for low-cost adaptation of the foundation model to specific scenarios. We reiterate that in real-life applications, the *test set* is inaccessible. Hence, we use our benchmarking strategy and select five high-ranking datasets on the target *training set* to finetune the model for 5 epochs. We perform

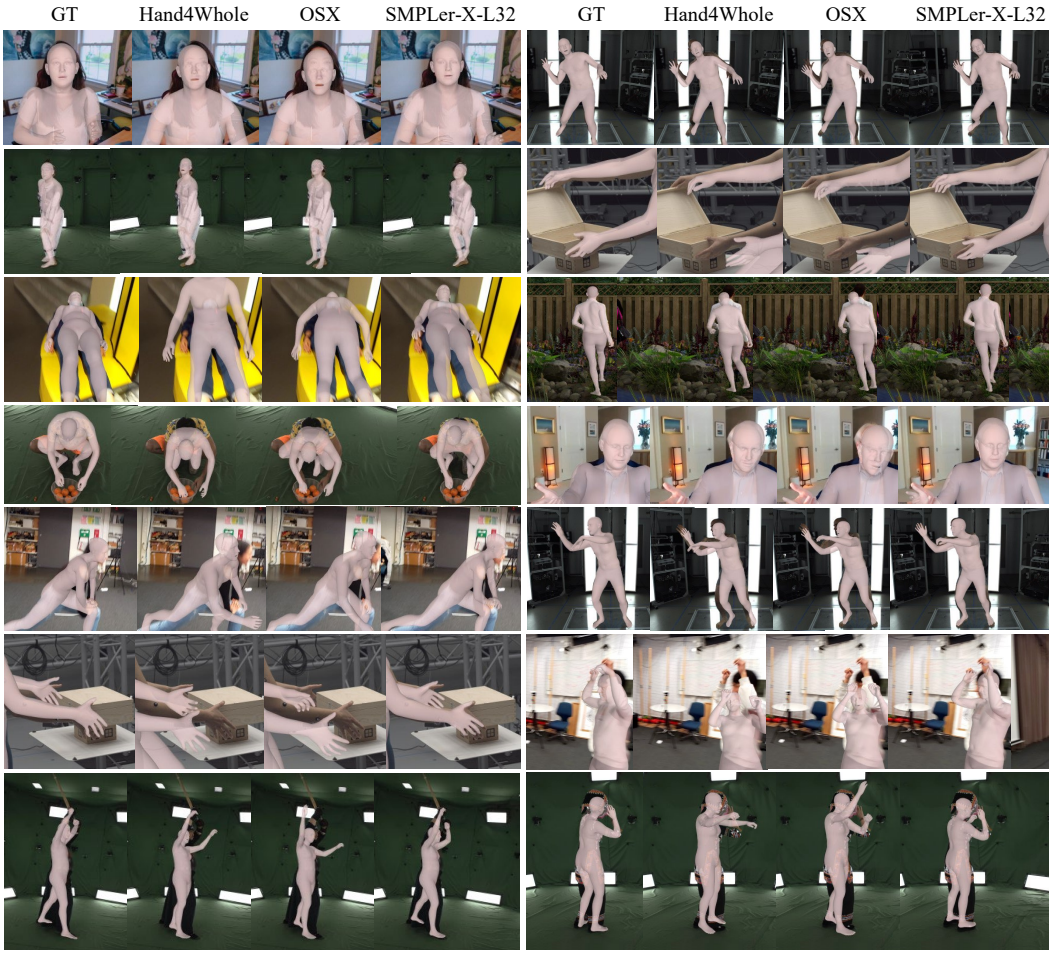

Figure 5: **Visualization.** We compare SMPLer-X-L32 with OSX [39] and Hand4Whole [46] (trained with the MSCOCO, MPII, and Human3.6M) in various scenarios such as those with heavy truncation, hard poses, and rare camera angles.

finetune experiments on ViT-L to match the backbone of current SOTA [39]. The results are shown in the same tables as the foundation models (Table 3, 4, 5, 6, 7, and 10), where finetuning always lead to substantial performance enhancement on the foundation models.

5 Conclusion

In this work, we benchmark datasets for EHPS that provide us insights for training and finetuning a foundation model. Our work is useful in three ways. First, our pretrained model (especially the backbone) can be a plug-and-play component of a larger system for EHPS and beyond. Second, our benchmark serves to gauge the performances of future generalization studies. Third, our benchmarking-finetuning paradigm can be useful for the rapid adaptation of any foundation model to specific scenarios. Specifically, users may collect a training set, evaluate pretrained models of various other datasets on it, and select the most relevant datasets to finetune a foundation model.

Limitations. First, although we use five comprehensive benchmark datasets to gauge the generalization capability, they may still be insufficient to represent the real-world distribution. Second, our experiments do not fully investigate the impact of various model architectures due to the prohibitive cost of training the foundation model.

Potential negative societal impact. As we study training strong EHPS models and release the pretrained models, they may be used for unwarranted surveillance or privacy violation.

Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). The project is also supported by NTU NAP and Singapore MOE AcRF Tier 2 (MOET2EP20221-0012).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[2] 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.

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022.

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022.

[5] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023.

[6] 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 *European conference on computer vision*, pages 561–578. Springer, 2016.

[7] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022.

[8] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021.

[9] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023.

[10] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer, 2020.

[11] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17840–17852, 2023.

[12] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. `https://github.com/open-mmlab/mmhuman3d`, 2021.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023.

[15] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021.

[16] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020.

[17] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1343–1351, 2021.

[18] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9919–9928, 2021.

[19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[20] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud sequences. *Advances in Neural Information Processing Systems*, 34:27940–27951, 2021.

[21] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022.

[22] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022.

[23] 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.

[24] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, pages 1–11. British Machine Vision Association, 2010.

[25] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011.

[26] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021.

[27] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. *3DV*, 2022.

[28] 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*, pages 7122–7131, 2018.

[29] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

[31] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. *arXiv preprint arXiv:2104.08527*, 2021.

[32] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021.

[33] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

[34] 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*, pages 6050–6059, 2017.

[35] 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*, 2023.

[36] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[37] 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 *CVPR*, pages 3383–3393. Computer Vision Foundation / IEEE, 2021.

[38] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022.

[39] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023.

[40] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[41] Qihao Liu, Adam Kortylewski, and Alan L Yuille. Poseexaminer: Automated testing of out-of-distribution robustness in human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2023.

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[43] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[44] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[45] Dushyant Mehta, Oleksandr Sotnychenko, F. Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. *2018 International Conference on 3D Vision (3DV)*, pages 120–130, 2018.

[46] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317, 2022.

[47] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022.

[48] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021.

[49] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[50] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021.

[51] 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*, pages 10975–10985, 2019.

[52] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5340–5348, 2019.

[53] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021.

[54] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2956–2966, 2023.

[55] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, September 2020.

[56] Yu Sun, Tianyu Huang, Qian Bao, Wu Liu, Wenpeng Gao, and Yili Fu. Learning monocular mesh recovery of multiple body parts via synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2669–2673. IEEE, 2022.

[57] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21970–21982, 2023.

[58] 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)*, pages 601–617, 2018.

[59] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. *arXiv preprint arXiv:2303.13796*, 2023.

[60] Yanjun Wang, Qingping Sun, Wenjia Wang, Jun Ling, Zhongang Cai, Rong Xie, and Li Song. Learning dense uv completion for human mesh recovery. *arXiv preprint arXiv:2307.11074*, 2023.

[61] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019.

[62] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[63] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. *arXiv preprint arXiv:2303.17368*, 2023.

[64] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, June 2023.

[65] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[66] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[67] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023.

[68] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 180–200. Springer, 2022.

[69] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019.

[70] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4822, 2021.

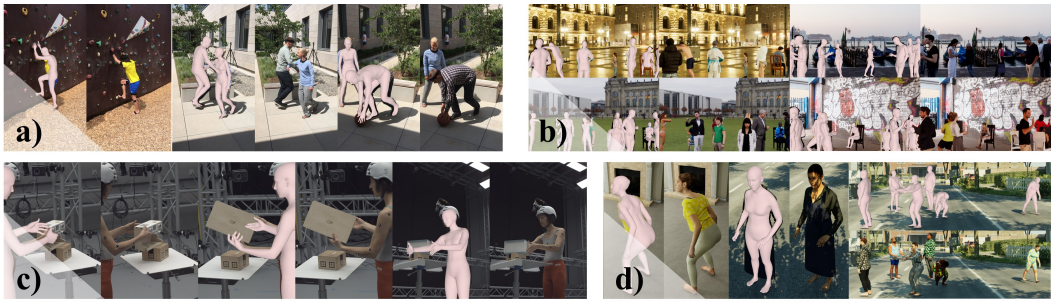

Figure 6: **Visualization** of dataset images and ground truth annotation. a) 3DPW. b) AGORA. c) ARCTIC. d) BEDLAM.

A Overview

Due to space constraints in the main paper, we elaborate the following here: additional details of the 32 datasets, including useful links to find their license statements and other ethics concerns in Sec. B; additional details of the architecture, training and finetuning stages in Sec. C; additional experiments and analyses on dataset distributions, training schemes, finetuning strategies, sampling strategies, and training domains in Sec. D; individual dataset ranking on the training sets of key evaluation benchmarks, and complete results of the foundation models on evaluation benchmarks in Sec. E.

B Additional Details of Datasets

B.1 Dataset Descriptions

This section describes the 32 datasets we study. Note that all these are public academic datasets, each holding a license. We follow the common practice to use them in our non-commercial research and refer readers to their homepages or papers for more details regarding *licenses* and their policies to ensure personal information protection.

3DPW [58] (Fig. 6a) is the first in-the-wild dataset with a considerable amount of data, captured with a moving phone camera and IMU sensors. It features accurate SMPL annotations and 60 video sequences captured in diverse environments. We follow the official definition of train, val, and test splits. Homepage: `https://virtualhumans.mpi-inf.mpg.de/3DPW/`.

AGORA [50] (Fig. 6b) is a synthetic dataset, rendered with high-quality human scans and realistic 3D scenes. It consists of 4240 textured human scans with diverse poses and appearances, each fitted with accurate SMPL-X annotations. There are 14K training images and 3K test images, and 173K instances. Homepage: `https://agora.is.tue.mpg.de/index.html`

ARCTIC [14] (Fig. 6c) is a lab-based hand-object interaction dataset. It features 10 subjects manipulating 11 objects. There are 210K frames of video sequences captured from 8 static cameras and one egocentric camera. Each frame is fitted with accurate SMPL-X annotations. We exclude the egocentric frames in our training as they only capture hands, and use 153.9K images in training. Homepage: `https://arctic.is.tue.mpg.de/`

BEDLAM [5] (Fig. 6d) is a synthetic dataset that includes a wide range of variations in terms of body shapes, motions, skin tones, hair, and clothing. It is created by combining 271 different body models, 27 hairstyles, and 111 types of clothing. The dataset includes 1691 clothing textures and 2311 human motions set in 95 HDRI and 8 3D scenes. Each scene typically consists of 1 to 10 people and offers diverse camera poses. Homepage: `https://bedlam.is.tue.mpg.de/index.html`

BEHAVE [4] (Fig. 7a) is a body human-object interaction dataset with multi-view RGB-D frames, SMPL-H parameters, object fits, and contacts information. BEHAVE includes about 15k frames in 5 locations with 8 subjects performing a range of interactions with 20 common objects. Homepage: `https://github.com/xiexh20/behave-dataset`.

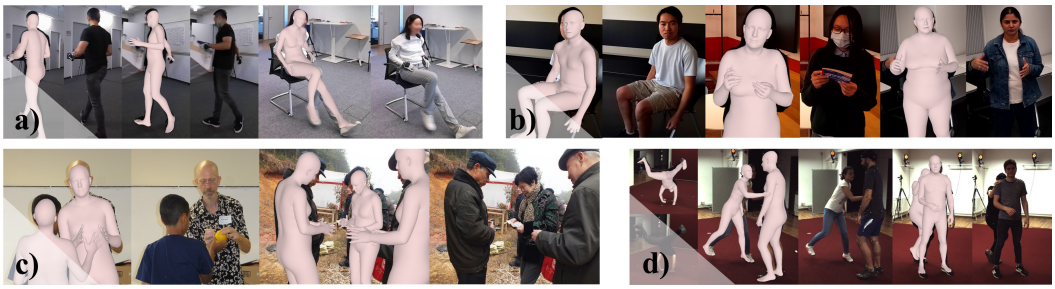

Figure 7: **Visualization** of dataset images and ground truth annotation. a) BEHAVE. b) EgoBody (EgoSet). c) CrowdPose. d) CHI3D.

CHI3D [16] (Fig. 7d) is a studio-based 3D motion capture dataset (Vicon) under multiple interaction scenarios, which includes 631 multi-view sequences with 2,525 contact events and 728,664 ground truth instances of 3D poses annotated with SMPL-X parameters. We use the open-source train set. Homepage: `https://ci3d.imar.ro`.

CrowdPose [36] (Fig. 7c) is an in-the-wild dataset focused on crowded cases. It contains 20K images in total and 80K human instances. In this paper, we use the annotations generated by NeuralAnnot [47], which fits the SMPL to the GT 2D joints and includes a total of ~35.7K annotated data. Homepage: `https://github.com/Jeff-sjtu/CrowdPose`

EgoBody [68] is a large-scale dataset that features 3D human motions and interaction with scenes. The data is captured by a multi-view rig for third-person view (MVSet, in Fig. 8a) and a head-mounted device for egocentric view (EgoSet, in Fig. 7b). The dataset consists of 125 sequences, 36 subjects, and 15 indoor scenes. We follow the official splits of training and test sets. Homepage: `https://sanweiliti.github.io/egobody/egobody.html`.

EHF [51] (Fig. 8b) contains 100 curated frames of one subject in an indoor studio setup. It provides SMPL-X aligned 3D mesh as the ground truth that accurately reflects the subject' diverse body, hand, and face articulations. It is usually used as a test set. The images are captured from a single camera. It is published along with SMPL-X. Homepage: `https://smpl-x.is.tue.mpg.de/index.html`.

FIT3D [18] (Fig. 8c) is a studio-based 3D motion capture dataset including 611 multi-view sequences with 2,964,236 images and corresponding ground truth instances of 3D shapes and poses annotated with SMPL-X parameters. Motion clips include 37 repeated exercises. We use the open-source train set. Homepage: `https://fit3d.imar.ro/`.

GTA-Human II (Fig. 8d) is an extended version of GTA-Human [8], a large-scale synthetic 3D single-human dataset generated with the GTA-V game engine, which features diversity. GTA-Human provides more than 1.4M of SMPL annotations in single-person scenes. In comparison, GTA-Human II includes multi-human scenarios with SMPL-X ground truth, obtained through SMPLify-X [51], which estimates SMPL-X parameters from ground truth keypoints collected in-game. The toolchain is provided by MMHuman3D [12]. The extended version contains 1.8M SMPL-X instances. Images are captured in 4K multi-person sequences, with about 600 subjects in different shapes and clothing, performing 20K daily human activity motion clips in six distinct categories of backgrounds, captured by camera angles in realistic distributions. Homepage: `https://caizhongang.github.io/projects/GTA-Human/`.

Human3.6M [23] (Fig. 9a) is a studio-based 3D motion capture dataset including 3.6M human poses and corresponding images captured by a high-speed motion capture system. In this paper, we use the annotation generated by NeuralAnnot [47], which fits the SMPL-X to the GT 2D joints and includes a total of ~312.2K annotated data. Homepage: `http://vision.imar.ro/human3.6m/description.php`

HumanSC3D [17] (Fig. 9b) is a studio-based 3D motion capture dataset including 1,032 multiple-view sequences featuring 5K contact events and 1.2M ground truth instances of 3D poses annotated with SMPL-X parameters. We use the open-source train set. Homepage: `https://sc3d.imar.ro/`.

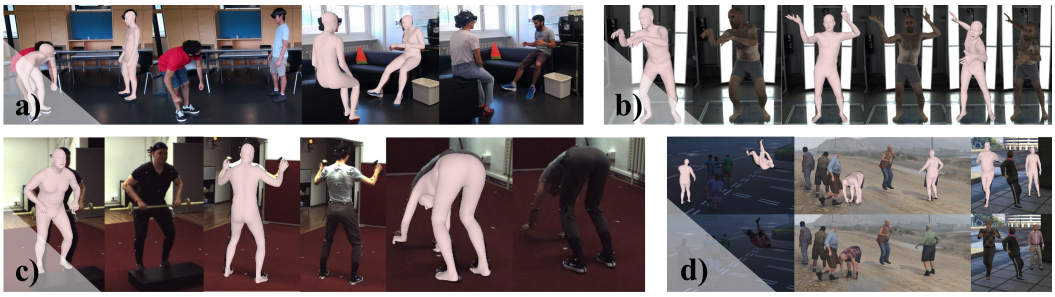

Figure 8: **Visualization** of dataset images and ground truth annotation. a) EgoBody (MVSet). b) EHF. c) FIT3D. d) GTA-Human.

Figure 9: **Visualization** of dataset images and ground truth annotation. a) Human3.6M. b) HumanSC3D. c) InstaVariety. d) LSPET.

InstaVariety [29] (Fig. 9c) is an in-the-wild dataset, containing 2.1M images collected from Instagram using 84 hashtags. We use the annotation generated by NeuralAnnot [47], which fits the SMPL to the GT 2D joints and includes a total of ~218.5K annotated data. Homepage: `https://github.com/akanazawa/human_dynamics/blob/master/doc/insta_variety.md`

LSPET [24] (Fig. 9d) is an in-the-wild dataset, and it contains 10K images. In this paper, we use the annotation generated by EFT [27], which fits the SMPL to the GT 2D joints and includes a total of 2,946 annotated data. Homepage: `http://sam.johnson.io/research/lspet.html`.

MPI-INF-3DHP [44] ((Fig. 10a) is captured with a multi-camera markerless motion capture system in constrained indoor and complex outdoor scenes. It records 8 actors performing 8 activities from 14 camera views. We use the annotations generated by NeuralAnnot [47], which fits the SMPL-X to the GT 2D joints and includes a total of 939,847 annotated data. Homepage: `https://vcai.mpi-inf.mpg.de/3dhp-dataset/`

MPII [2] ((Fig. 10b) is a widely used in-the-wild dataset that offers a diverse collection of approximately 25K images. Each image within the dataset contains one or more instances, resulting in a total of over 40K annotated people instances. Among the 40K samples, ~28K samples are used for training, while the remaining samples are reserved for testing. We use the annotations generated by NeuralAnnot [47], which fits the SMPL-X to the GT 2D joints and includes a total of ~28.9K annotated data. Homepage: `http://human-pose.mpi-inf.mpg.de/`

MSCOCO [40] (Fig. 10c) is a large-scale object detection, segmentation, keypoint detection, and captioning dataset. The subset for the keypoint detection contains more than 200K images and 250K person instances. We use the annotations generated by NeuralAnnot [47], which fits the SMPL-X to the GT 2D joints and includes a total of ~149.8K annotated data. Homepage: `https://cocodataset.org/#home`

MTP [48] (Fig. 10d) is an in-door dataset containing images of actors mimicking different hard SMPL-X poses with self-contact. There are 3.7K images from 148 subjects with pseudo ground-truth SMPL-X parameters and 2D keypoints. We use 3.2K instances in training. Homepage: `https://tuch.is.tue.mpg.de/`

Figure 10: **Visualization** of dataset images and ground truth annotation. a) MPI-INF-3DHP. b) MPII. c) MSCOCO. d) MTP.

Figure 11: **Visualization** of dataset images and ground truth annotation. a) MuCo-3DHP. b) OCHuman. c) PoseTrack. d) PROX.

MuCo-3DHP [45] (Fig. 11a) is an in-door multi-person dataset composited by cropping and overlaying person in MPI-INF-3DHP[44] with segmentation masks. It has 400K frames and contains 8 subjects with 2 different clothing for each subject. It is shot with 12 different camera positions. It has ground truth 3D keypoints and fitted SMPL parameters. We use 465.3K annotated data in training. Homepage: `https://vcai.mpi-inf.mpg.de/projects/SingleShotMultiPerson/`.

OCHuman [69] (Fig. 11b) is an in-the-wild datset, and it focuses on heavily occluded human. This dataset contains 8,110 detailed annotated human instances within 4,731 images. We use the annotations generated by EFT [27], which fits the SMPL to the GT 2D joints and includes a total of 2,495 annotated data. Homepage: `https://github.com/liruilong940607/OCHumanApi`

PoseTrack [1] (Fig. 11c) is a large-scale benchmark for multi-person pose estimation and tracking in videos. It contains 514 videos and includes 66,374 frames. We use the annotations generated by EFT [27], which fits the SMPL to the GT 2D joints and includes a total of ~28.5K annotated data. Homepage: `https://posetrack.net`

PROX [19] (Fig. 11d) qualitative dataset is a human-scene interaction dataset that showcases 12 indoor scenes and 20 subjects engaging with these scenes. It comprises 100K RGB-D frames with pseudo-ground-truth SMPL-X fittings. During training, only the RGB images are utilized, and they are horizontally flipped to align with the SMPL-X annotations. We use 88.1K instances for training. Homepage: `https://prox.is.tue.mpg.de/`.

DNA-Rendering [9] (Fig. 12a) is a large-scale multi-view studio-based dataset with different resolutions (main set and HiRes set) that features diversity in motion, clothing, and object interactions. DNA-Rendering has more than 1.5K human instances and 5K motion sequences with up to 60 RGB views and 4 Kinect views. Corresponding SMPL-X annotation is based on HuMMan [7]. We separate the 60 RGB views into 48 and 12 views based on different camera distributions and captured resolutions. Homepage: `https://dna-rendering.github.io/`.

SPEC [32] (Fig. 12c) is a synthetic dataset featuring diverse and unique camera viewpoints. It has 22,191 images with 71,982 ground truth instances with SMPL parameters as a train set and 3,783

Figure 12: **Visualization** of dataset images and ground truth annotation. a) DNA-Rendering. b) SSP3D. c) SPEC. d) RICH.

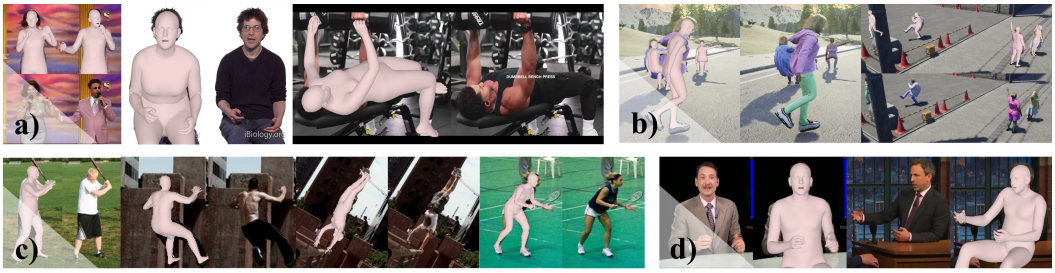

Figure 13: **Visualization** of dataset images and ground truth annotation. a) UBody. b) SynBody. c) UP3D. d) Talkshow.

images with 12,071 ground truth instances as the test set. Homepage: `https://spec.is.tue.mpg.de/index.html`.

RICH [22] (Fig. 12d) is a human-scene contact dataset. It includes a comprehensive collection of 142 single or multi-person multiview videos capturing 22 subjects in 5 static indoor or outdoor scenes with 6-8 static cameras. RICH comprises a rich set of resources, including a total of 90K posed 3D body meshes, each associated with dense full-body contact labels in both SMPL-X and SMPL mesh topology. We convert the original image from .png and .bmp to .jpg and train the model with the train set, which includes ~243.4K instances. Homepage: `https://rich.is.tue.mpg.de/index.html`

SSP3D [55] SSP-3D (Fig. 12b) is a small-scale dataset consisting of 311 images of persons in tight-fitted clothes in sports, with a variety of body shapes and poses. Pseudo-ground-truth SMPL body model parameters obtained via multi-frame optimization with shape consistency. Homepage: `https://github.com/akashsengupta1997/SSP-3D`.

SynBody [63] (Fig. 13b) is a large-scale synthetic dataset featuring a massive number of diverse subjects and high-accuracy annotations which includes multi-person image instances with 3D pose and shape annotations. SynBody covers 10K human body models, 1K actions, and many viewpoints. Annotations include both accurate SMPL and SMPL-X parameters. Synbody also features layered human body models and clothes. We sample a set with ~600K instances in our study. Homepage: `https://maoxie.github.io/SynBody/`.

Talkshow [64] (Fig. 13d) is a large-scale dataset featuring talking videos of 4 subjects in 4 different scenarios. It contains 26.9 hours of video clips at 30 FPS and has synchronized audio and fitted SMPL-X annotations. We obtain the video clips from the author and convert them to images, including of 332.7K instances. Homepage: `https://talkshow.is.tue.mpg.de/`.

UBody [39] (Fig. 13b) is a large-scale dataset that features a diverse range of real-life scenarios that cater to various downstream tasks, such as fitness videos, VLOGs, movies, online classes, video conferences, talk shows, and sign languages. In these scenarios, typically only the subject's upper body is visible. Heavy truncation and a focus on expressive gestures and facial expressions make UBody especially challenging. Homepage: `https://github.com/IDEA-Research/OSX`.

UP3D [34] (Fig. 13c) is an in-the-wild dataset containing 7,126 images. To obtain 3D high-quality annotations, it extends the SMPLify [6] and fits a pseudo label (SMPL) for each image. Homepage: `https://files.is.tuebingen.mpg.de/classner/up/`

C Additional Details of Foundation Model

C.1 Architecture

SMPLer-X utilizes a minimalistic design. Before entering the backbone, the image is cropped by a whole body bounding box and resized to I with (height, width) as (512, 384). The image is then tokenized into 32×24 patches with patch size 16, and undergoes patch embedding, and positional encoding is added to obtain image tokens T_{img}. T_{task} is additional learnable tokens (task tokens) that are concatenated with T_{img}. The tokens are processed with *backbone* (denoted as ViT). Leveraging the scalability of ViT [13], we are able to experiment with various model sizes. In the *neck*, the processed image tokens, $T^{'}_{img}$ are used to predict face and hand bounding boxes. The predicted bounding boxes are used in the ROI (regions of interest) module to crop features from $T^{'}_{img}$, which is re-organized and undergoes transposed convolution (deconv), and fed into hand and face heads. The body head takes in both $T^{'}_{img}$ (omitted in the illustration) and $T^{'}_{task}$. The hand and body heads consist of a positional module to predict 3D keypoints, and a regressor module to predict parameters, whereas for the face head, we follow OSX [39] to include only a regressor module. We highlight that training foundation models are very expensive. Hence, we do not conduct extensive architectural searches in our study. We use SMPLer-X as a simple baseline with the essential components, which (*e.g.*, backbone) can be directly used in future research. In addition, the data selection strategies in our study are likely to be applicable to any other architectures.

C.2 Training Details

The training is conducted on 16 V100 GPUs, with a total batch size of 512 (256 for ViT-Huge) for 10 epochs. Specifically, SMPLer-X-L20 takes more than 400 GPU hours to train and SMPLer-X-H32 takes more than 700 GPU hours to train. We use Adam optimizer with cosine annealing for both training and fine-tuning. The learning rate for training is 1×10^{-5} with the minimum learning rate set to 1×10^{-6}, while the learning rate for finetuning is 1×10^{-5} with the minimum learning rate set to 5×10^{-7}.

C.3 Adaption of SMPL/SMPL-X Annotations.

While we strive to utilize as many datasets as possible in our study, we find that there are only a few datasets with neutral SMPL-X annotations and many datasets with female/male (gendered) SMPL-X annotations or SMPL annotations. An intuitive solution is to use the official fitting tool [51], however, this optimization-based tool is relatively slow to convert a large number of annotations (fitting takes 241±126 seconds per frame). Hence, we experiment with a new approach.

For gendered SMPL-X annotations, we train a small adapter network A (consisting of three layers of fully-connected layers) that takes in gendered body shape parameters $\beta_{f/m}$ and converts it neutral body shape parameters such that the following loss is minimized:

$$\mathcal{L} = ||M_{f/m}(\theta, \beta_{f/m}) - M_n(\theta, A(\beta_{f/m})||_2 \tag{1}$$

where $M_{f/m}$ are gendered SMPL-X body model, and M_n is the neutral SMPL-X body model, θ is body pose is obtained by random sampling in the embedding space of VPoser [51]. We test our adapter on AGORA [50] and find that the vertex-to-vertex error between ground truth gendered SMPL-X mesh and neutral SMPL-X with adapted neutral β is 8.4 mm, which we consider to be sufficiently small. This approach is very fast (0.09 seconds per frame). Hence, we apply our adapter on AGORA, EgoBody, DNA-Rendering, and RICH.

However, we empirically find that the adapter does not work well across significantly different topologies (*i.e.* SMPL and SMPL-X), training similar adapters results in a 27.1 mm vertex-to-vertex error. Hence, for datasets with SMPL annotations, we only supervise ground truth global orientation and body pose. Although this is a slight abuse of the parameters (SMPL and SMPL-X parameters are not directly transferable), we find in our experiments that such a strategy leads to performance gains.

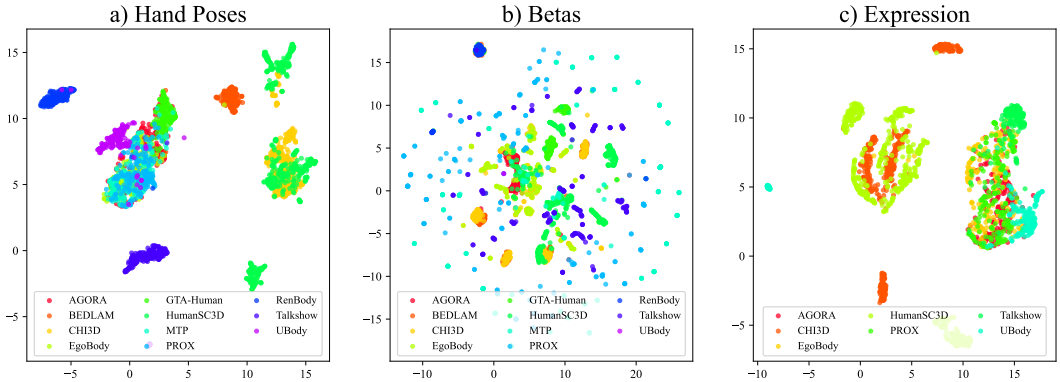

Figure 14: **Comparisons of hand pose, shape beta, and facial expression parameters distributions among different datasets**. We illustrate these distributions with UMAP [43]. The two axes are the two dimensions of the embedded space and have no unit.

Table 11: **Training schemes**. We study the different training schemes by comparing the model trained with the Top 5 / Top 10 datasets with the Bottom 5 / Bottom 10 datasets according to our individual dataset benchmark rankings.

Method	Dataset	#Instance	MPE(mm)
SMPLer-X-B	Top5	0.75M	103.47
SMPLer-X-B	Bottom5	0.75M	115.61
SMPLer-X-B	Top10	1.5M	89.20
SMPLer-X-B	Bottom10	1.5M	115.10

D Additional Experiments and Analyses

D.1 More Distribution Comparisons

In Fig. 14, we plot more distributions of additional parameters: a) hand poses, b) betas (body shape), and c) facial expression, all via UMAP dimension reduction. Datasets without proper SMPL-X parameters (*e.g.*, SMPL annotation only, or pseudo-annotated that typically have invalid hand poses) are not included in the study. For hand poses, we concatenate both left and right-hand parameters in rotation matrix representation. For betas and expression, we directly use their first 10 components. It is observed that datasets such as DNA-Rendering, CHI3D, HumanSC3D, and Talkshow form distinct clusters for hand poses and betas, and it is difficult to find any dataset to provide a well-spread coverage. For expression, there is still a lack of diverse datasets.

D.2 Training Schemes

As shown in Table 11, we perform the ablation study for the training scheme. We investigate the effect of dataset selection. We selecte the bottom 5 and bottom 10 datasets according to our individual dataset benchmark rankings and trained the SMPLer-X-B model with the same number of instances as used in training with the top 5 and top 10 datasets.

It is proved that our training scheme is efficient. Selecting the top 5 or top 10 datasets according to the single dataset benchmark leads to a much better performance compared to selecting the bottom 5 or bottom 10 datasets. The foundation model can benefit from adding higher-ranked (i.e., Top 5/10) data into training, while lower-ranked data (i.e., Bottom 5/10) is not as effective in improving the model's performance. Despite this, we finetune the entire network in all other finetuning experiments.

D.3 Finetuning Strategies

In Table 12, we evaluate different strategies that finetune different parts of our foundation model. We observe that finetuning only the neck and head is very efficient: it achieves even slightly better perfor-

Table 12: **Finetuning strategies**. We study the different finetuning strategies by freezing the parameters in different parts of the network. Models are tested on UBody test set, and † denotes the models that are finetuned on UBody train set.

			PA-PVE (*mm*)			PVE (*mm*)		
Method	Finetune	#Param.	All	Hands	Face	All	Hands	Face
SMPLer-X-H32	-	662M	29.9	9.8	2.6	54.5	36.4	20.6
SMPLer-X-H32†	Full network	662M	27.8	9.0	2.3	51.3	32.6	19.1
SMPLer-X-H32†	Neck+Head	31M	27.8	9.0	2.3	51.1	32.5	19.1
SMPLer-X-H32†	Head	5M	29.9	9.7	2.6	54.2	35.9	20.6

Table 13: **Data sampling strategies**. We trained SMPLer-X-H32 models with different data sampling strategies.

Method	Strategy	#Instance	MPE(mm)
SMPLer-X-H32	balanced	4.5M	63.08
SMPLer-X-H32	weighted	4.5M	62.12
SMPLer-X-H32	concat	5.6M	63.32

mance than finetuning the entire network, with much fewer learnable parameters. We speculate that after training with a large number of datasets, the backbone is already very strong and generalizable. Hence, finetuning the backbone does not yield much performance improvement.

D.4 Data Sampling Strategies

As for the sampling strategy, we did the ablation study on three different strategies, including 1) Balanced: we set all the datasets to have the same length; 2) Weighted: we set the dataset length according to the individual dataset benchmark rankings. Specifically, we sort the datasets based on their rankings and then assign weights to each dataset. As a result of this weighting, the datasets are upsampled or downsampled so that the lengths of the datasets are adjusted to an arithmetic sequence. The length of the dataset with the highest ranking is 4 times that of the dataset with the lowest ranking, and the sum of the total lengths of all datasets is fixed; 3) Concat: we simply concatenate all the datasets with their original length.

The performance of the foundation model is not sensitive to the sampling strategy as shown in Table 13, while the balanced strategy is more intuitive, easy to implement, and efficient, the weighted strategy may have more potential with more effort in weight tuning.

Table 14: **Impact of training domains.** We investigate the impact of seeing the train split of a benchmark dataset during training and how this may affect the generalizability of a model. MPE: mean primary error of AGORA-val, EgoBody-EgoSet, UBody, 3DPW, and EHF. The yellow shaded numbers denote that the corresponding train split is used in training. Top-1 values are bolded, and the second best values are underlined. Except for 3DPW using MPJPE as the metric, other datasets are evaluated via PVE. Unit: mm. #Data.: number of datasets used in the training. #Seen: number of evaluation benchmarks' used in the training, note here that only benchmarks that are included in the MPE computation are counted, thus excluding ARCTIC and DNA-Rendering-HiRes. *: not following the standard dataset selection scheme.

#Data.	#Seen	Model	MPE	AGORA [50]	EgoBody [68]	UBody [39]	3DPW [58]	EHF [55]	ARCTIC [14]	DNA-R-HiRes [9]
5	1	SMPLer-X-L5	100.8	89.1	101.6	114.0	102.8	96.7	99.8	90.6
5	4	SMPLer-X-L*	85.2	96.7	81.9	68.1	95.5	83.6	103.6	98.2
10	2	SMPLer-X-L10	80.6	82.6	69.7	104.0	82.5	64.0	76.9	76.2
10	4	SMPLer-X-L*	72.7	84.0	71.6	62.8	81.7	63.4	80.8	75.8
20	4	SMPLer-X-L20	70.5	80.7	66.6	61.5	78.3	65.4	52.2	77.7
32	4	SMPLer-X-L32	**66.2**	**74.2**	**62.2**	**57.3**	**75.2**	**62.3**	**48.6**	**54.4**

Table 15: **AGORA Val set**. † denotes methods that are finetuned on the AGORA training set.

Method	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [46]†	73.2	9.7	4.7	183.9	72.8	81.6
OSX [39]∗	45.0	8.5	3.9	79.6	48.2	37.9
SMPLer-X-S5	72.1	10.2	5.1	119.0	66.8	58.9
SMPLer-X-S10	67.5	10.2	4.8	116.0	65.2	57.5
SMPLer-X-S20	62.1	10.0	4.4	109.2	63.3	55.2
SMPLer-X-S32	58.7	9.8	4.2	105.2	61.9	53.9
SMPLer-X-B5	63.8	9.6	4.8	102.7	59.0	50.8
SMPLer-X-B10	58.4	9.5	4.6	97.8	57.8	49.1
SMPLer-X-B20	56.9	9.3	4.3	95.6	56.5	47.9
SMPLer-X-B32	52.0	9.2	4.1	88.0	54.5	45.9
SMPLer-X-L5	56.1	9.2	4.3	88.3	53.0	43.3
SMPLer-X-L10	50.6	9.1	4.1	82.6	51.9	42.3
SMPLer-X-L20	48.6	8.9	4.0	80.7	51.0	41.3
SMPLer-X-L32	45.1	8.7	3.8	74.2	47.8	38.7
SMPLer-X-H5	57.8	9.1	4.2	89.0	52.6	42.6
SMPLer-X-H10	51.0	9.0	4.0	81.4	51.4	40.5
SMPLer-X-H20	47.1	8.8	3.9	77.5	49.5	39.4
SMPLer-X-H32	42.9	8.5	**3.7**	69.5	45.6	35.9
SMPLer-X-H32†	**41.0**	**8.2**	**3.7**	**65.4**	**43.8**	**34.0**

Table 16: **EHF**. As EHF does not have a training set, we do not perform finetuning.

Method	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	50.3	**10.8**	5.8	76.8	**39.8**	26.1
OSX [39]	48.7	15.9	6.0	70.8	53.7	26.4
SMPLer-X-S5	70.7	16.0	5.9	100.5	64.0	27.1
SMPLer-X-S10	60.5	16.0	5.7	89.9	59.1	22.3
SMPLer-X-S20	51.0	15.5	5.5	86.6	54.7	22.1
SMPLer-X-S32	50.5	14.8	5.2	74.1	54.6	20.0
SMPLer-X-B5	61.4	15.4	5.8	96.1	58.4	27.1
SMPLer-X-B10	46.7	15.7	5.6	74.7	55.1	21.3
SMPLer-X-B20	41.9	15.9	5.3	73.0	53.7	20.8
SMPLer-X-B32	40.7	14.5	5.2	67.3	52.1	20.6
SMPLer-X-L5	53.9	14.7	5.9	89.5	57.8	29.9
SMPLer-X-L10	40.7	15.6	5.3	64.0	52.9	18.1
SMPLer-X-L20	37.8	15.0	5.1	65.4	49.4	17.4
SMPLer-X-L32	**37.1**	14.1	**5.0**	62.4	47.1	**17.0**
SMPLer-X-H5	47.0	14.3	5.9	68.3	55.6	25.0
SMPLer-X-H10	40.1	15.6	5.2	**56.6**	50.2	18.9
SMPLer-X-H20	39.0	14.4	**5.0**	59.4	47.1	17.8
SMPLer-X-H32	39.0	14.8	**5.0**	56.8	42.2	19.0

D.5 Training Domains

In Table 14, we further study the impact of training domains. It is clear that in-domain training (including the training split of a dataset in the training, and testing on the test split of the same dataset) is highly effective, as "seeing" the dataset always brings significant performance improvement. However, we highlight that having out-of-domain training sets in the training is also highly effective: with 4 seen datasets fixed, SMPLer-X benefits tremendously from having 10, 20, and 32 datasets in training in terms of MPE. It is worth noting that training with a lot of datasets especially benefits out-of-domain ("unseen" benchmarks) performance as errors on EHF, ARCTIC, and DNA-Rendering-HiRes decrease with more datasets in the training set. Lastly, training on 32 datasets with our SMPLer-X-L obtains the best performance with 66.2 mm MPE, making it a strong and effective SMPL-X estimator.

E Complete Results

E.1 Benchmarking EHPS Datasets on Training Sets

In the main paper, we benchmark individual datasets on the testing sets of the key EHPS evaluation benchmarks. However, this dataset benchmark is unsuitable for selecting top datasets for training EHPS, as the ranking leaks information about the testing sets to some extent. Hence, we construct a new benchmark that ranks EHPS datasets on the training set of AGORA, UBody, EgoBody, and 3DPW (EHF is omitted as it does not have a training set) in Table 22.

E.2 Complete Results of Foundation Models on Evaluation Benchmarks

We show complete results including our strongest foundation model SMPLer-X-H32 on AGORA validation set (Table 15), UBody (Table 17), EgoBody-EgoSet (Table 18), EHF (Table 16), ARCTIC (Table 19) and DNA-Rendering-HiRes (Table 20).

Table 17: **UBody.** † denotes the methods that are finetuned on the UBody training set.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	42.2	**8.3**	3.1	95.7	39.0	31.2
OSX [39]†	42.2	8.6	**2.0**	81.9	41.5	21.2
SMPLer-X-S5	53.9	11.8	3.9	110.1	59.4	34.5
SMPLer-X-S10	50.4	11.5	3.7	107.7	57.4	32.8
SMPLer-X-S20	37.5	11.1	3.2	70.7	49.6	26.1
SMPLer-X-S32	36.4	10.7	3.0	68.1	47.8	25.0
SMPLer-X-B5	52.3	11.9	3.8	105.8	56.9	32.6
SMPLer-X-B10	49.7	12.0	3.6	107.3	57.1	31.7
SMPLer-X-B20	35.5	11.0	3.0	65.3	46.9	23.4
SMPLer-X-B32	33.7	10.8	2.8	63.3	43.9	22.7
SMPLer-X-L5	51.8	12.5	3.6	110.8	56.3	37.5
SMPLer-X-L10	48.0	12.8	3.5	104.0	56.1	32.0
SMPLer-X-L20	33.2	10.6	2.8	61.5	43.3	23.1
SMPLer-X-L32	30.9	10.2	2.7	57.3	39.2	21.6
SMPLer-X-H5	48.1	12.1	3.7	102.1	53.3	33.4
SMPLer-X-H10	48.5	12.6	3.5	100.7	54.8	30.9
SMPLer-X-H20	32.8	10.3	2.8	59.9	41.0	22.7
SMPLer-X-H32	29.9	9.8	2.6	54.5	36.4	20.6
SMPLer-X-H32†	**27.8**	9.0	2.3	**51.3**	**32.6**	**19.1**

Table 18: **EgoBody-EgoSet.** † are finetuned on the EgoBody-EgoSet training set.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	58.8	**9.7**	3.7	121.9	50.0	42.5
OSX [39]†	45.3	10.0	3.0	82.3	46.8	35.2
SMPLer-X-S5	62.8	10.8	4.1	114.2	53.3	44.3
SMPLer-X-S10	52.2	10.0	3.4	88.6	48.6	37.6
SMPLer-X-S20	48.1	10.0	3.3	84.3	47.2	37.8
SMPLer-X-S32	46.0	10.0	3.1	82.5	46.0	36.2
SMPLer-X-B5	59.4	10.6	4.0	108.1	48.0	40.0
SMPLer-X-B10	45.3	10.1	3.2	76.4	45.5	32.4
SMPLer-X-B20	43.8	9.9	3.2	75.5	44.6	32.7
SMPLer-X-B32	40.7	9.9	3.1	72.7	43.7	32.4
SMPLer-X-L5	52.9	10.5	3.8	98.7	45.2	39.1
SMPLer-X-L10	40.5	10.0	3.0	69.7	43.1	32.0
SMPLer-X-L20	38.9	9.9	3.0	66.6	42.7	31.8
SMPLer-X-L32	36.3	9.8	2.9	62.2	41.4	30.7
SMPLer-X-H5	48.0	10.5	3.4	87.4	43.5	37.5
SMPLer-X-H10	38.8	10.0	2.9	65.7	42.6	31.1
SMPLer-X-H20	36.7	9.8	2.9	63.5	41.3	30.8
SMPLer-X-H32	34.3	9.8	2.7	59.5	**39.6**	28.7
SMPLer-X-H32†	**33.9**	10.0	**2.5**	**57.0**	40.2	**27.1**

Table 19: **ARCTIC.** † denotes the methods that are finetuned on the ARCTIC training set.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	63.4	18.1	4.0	136.8	54.8	59.2
OSX [39]†	33.0	18.8	3.3	58.4	39.4	30.4
SMPLer-X-S5	66.1	**16.7**	4.0	117.3	58.7	46.5
SMPLer-X-S10	58.8	17.5	3.2	104.6	56.6	41.1
SMPLer-X-S20	37.6	18.9	2.7	58.7	45.2	30.5
SMPLer-X-S32	34.5	18.9	2.7	55.3	42.9	28.9
SMPLer-X-B5	66.3	16.9	3.4	105.4	55.6	41.4
SMPLer-X-B10	54.0	17.9	2.5	85.2	53.4	35.0
SMPLer-X-B20	34.9	18.9	2.7	56.3	40.9	29.6
SMPLer-X-B32	31.9	19.0	2.8	52.6	40.1	27.4
SMPLer-X-L5	57.2	17.0	2.9	95.1	52.8	37.7
SMPLer-X-L10	46.9	18.1	2.3	76.9	50.8	33.2
SMPLer-X-L20	31.9	18.9	2.5	52.2	39.3	27.0
SMPLer-X-L32	29.4	18.9	2.7	48.6	38.8	26.8
SMPLer-X-H5	49.3	17.4	2.5	79.9	49.3	33.9
SMPLer-X-H10	41.4	18.8	**2.1**	71.6	49.3	30.9
SMPLer-X-H20	29.3	18.9	2.5	48.5	38.3	26.3
SMPLer-X-H32	**27.6**	18.7	2.6	**44.6**	**36.9**	**24.6**
SMPLer-X-H32†	27.7	18.8	2.6	44.7	37.0	24.7

Table 20: **DNA-Rendering-HiRes**. † are finetuned on the DNA-Rendering-HiRes training set.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	62.8	11.0	4.2	111.4	56.4	52.6
OSX [39]†	43.5	7.5	3.5	67.1	43.3	38.2
SMPLer-X-S5	70.9	10.4	4.7	104.9	57.6	49.7
SMPLer-X-S10	63.9	11.0	4.4	98.4	57.0	47.3
SMPLer-X-S20	55.6	10.2	4.4	87.3	53.3	46.2
SMPLer-X-S32	47.1	7.7	3.5	70.1	46.9	39.0
SMPLer-X-B5	59.9	10.5	4.3	91.1	50.5	44.6
SMPLer-X-B10	53.3	11.5	4.1	83.7	50.9	42.4
SMPLer-X-B20	50.7	11.7	4.2	83.3	50.9	43.5
SMPLer-X-B32	40.9	7.4	3.4	61.9	40.5	36.6
SMPLer-X-L5	52.4	10.3	4.0	85.9	47.6	44.5
SMPLer-X-L10	47.0	11.2	3.8	76.2	47.8	41.7
SMPLer-X-L20	44.4	11.1	4.5	77.7	47.5	43.2
SMPLer-X-L32	35.8	7.2	3.2	54.4	36.7	34.0
SMPLer-X-H5	53.9	10.3	3.9	81.9	46.3	40.7
SMPLer-X-H10	47.4	10.9	3.7	76.2	47.0	39.0
SMPLer-X-H20	43.0	11.2	3.8	72.8	45.6	40.5
SMPLer-X-H32	34.0	**7.1**	**3.1**	51.4	34.5	32.0
SMPLer-X-H32†	**32.7**	**7.1**	**3.1**	**49.8**	**33.2**	**30.8**

Table 21: **3DPW.** †denotes the models that are finetuned on the 3DPW training set. Only whole-body (SMPL-X) methods are listed. Unit: *mm*.

Method	MPJPE	PA-MPJPE
Hand4Whole [46]	86.6	54.4
OSX [39]†	86.2	60.6
SMPLer-X-S5	110.2	79.1
SMPLer-X-S10	97.4	69.0
SMPLer-X-S20	87.4	60.0
SMPLer-X-S32	83.2	57.1
SMPLer-X-B5	104.8	72.0
SMPLer-X-B10	89.9	62.7
SMPLer-X-B20	83.5	57.6
SMPLer-X-B32	80.3	53.4
SMPLer-X-L5	97.8	62.6
SMPLer-X-L10	82.5	56.0
SMPLer-X-L20	78.3	52.1
SMPLer-X-L32	75.2	50.5
SMPLer-X-H5	88.3	60.3
SMPLer-X-H10	78.7	54.8
SMPLer-X-H20	74.4	50.9
SMPLer-X-H32	75.0	50.6
SMPLer-X-H32†	**71.7**	**48.0**

Table 22: **Selection of training datasets by ranking on the training set of key benchmarks.** For each dataset, we evaluate a model trained on the training set and on the *training* sets of four major benchmarks: AGORA, UBody, EgoBody (EgoSet), and 3DPW. Datasets are then ranked by MPE. ★: ranking on MPE. Top 1 values on each benchmark are bolded, and the rest of Top-5 are underlined.

Dataset	**MPE**↓	AGORA [50]↓	UBody [39]↓	EgoBody [68]↓	3DPW [58]↓
BEDLAM [5]	**124.7**	167.8	126.7	106.3	**98.1**
AGORA [50]	129.9	**131.7**	124.4	134.2	131.2
GTA-Human [8]	135.1	164.2	137.6	135.2	103.5
SynBody [63]	138.6	172.3	146.0	129.7	106.3
InstaVariety [29]	139.6	198.2	128.4	131.6	100.6
MSCOCO [40]	139.7	196.8	110.4	130.5	121.1
SPEC [32]	150.0	166.2	138.8	155.4	139.7
EgoBody-MVSet [68]	151.8	193.3	194.7	119.7	99.3
MPII [2]	152.0	205.5	127.3	143.3	131.9
RICH [22]	155.7	198.9	171.8	136.9	115.2
Egobody-EgoSet [68]	157.1	213.6	123.5	**63.6**	134.1
CrowdPose [36]	162.3	213.0	133.7	146.2	156.3
MuCo-3DHP [45]	163.4	193.2	189.7	151.1	119.7
UBody [39]	166.6	212.9	**61.5**	137.6	149.2
PROX [19]	167.3	205.1	186.8	145.2	132.1
MPI-INF-3DHP [44]	167.5	221.3	167.4	150.0	131.4
PoseTrack [1]	177.0	219.2	165.4	173.2	150.2
BEHAVE [3]	179.0	204.8	212.3	167.2	131.8
HumanSC3D [17]	184.8	213.8	237.7	174.8	112.9
CHI3D [16]	192.3	209.2	256.7	180.7	122.5
Human3.6M [23]	207.4	224.5	282.4	210.7	112.1
DNA-R.-HiRes [9]	207.5	231.1	275.4	189.4	134.0
ARCTIC [14]	222.5	303.6	205.9	177.3	203.2
Talkshow [64]	225.3	290.0	132.2	188.1	290.8
UP3D [34]	226.0	257.4	226.8	208.4	211.6
3DPW [58]	230.6	231.3	266.0	194.5	140.6
DNA-Rendering [9]	253.2	288.7	342.5	234.4	147.2
MTP [48]	270.5	272.8	284.8	259.2	265.4
FIT3D [18]	272.9	323.5	392.8	242.7	132.5
OCHuman [69]	282.3	307.7	266.7	261.5	293.4
LSPET [25]	330.2	361.6	301.8	317.3	340.2
SSP3D [55]	512.0	545.9	533.4	529.7	439.1