

SMPLest-X: Ultimate Scaling for Expressive Human Pose and Shape Estimation

Wanqi Yin¹, Zhongang Cai², *Member, IEEE*, Ruisi Wang, Ailing Zeng³, *Member, IEEE*, Chen Wei, Qingping Sun⁴, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang⁵, Lei Zhang⁶, *Fellow, IEEE*, Chen Change Loy⁷, *Senior Member, IEEE*, Atsushi Yamashita⁸, *Senior Member, IEEE*, Lei Yang⁹, and Ziwei Liu¹⁰, *Member, IEEE*

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 focus on training innovative architectural designs on confined datasets. In this work, we investigate the impact of scaling up EHPS towards a family of generalist foundation models. 1) *For data scaling*, we perform a systematic investigation on 40 EHPS datasets, encompassing 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. Ultimately, we achieve diminishing returns at 10 M training instances from diverse data sources. 2) *For model scaling*, we take advantage of vision transformers (up to ViT-Huge as the backbone) to study the scaling law of model sizes in EHPS. To exclude the influence of algorithmic design, we base our experiments on two minimalist architectures: SMPLer-X, which consists of an intermediate step for hand and face localization, and SMPLest-X, an even simpler version that reduces the network to its bare essentials and highlights significant advances in the capture of articulated hands. With Big Data and the large model, the foundation models exhibit strong performance across diverse test benchmarks and excellent transferability to even unseen environments. Moreover, our finetuning strategy turns the *generalist* into *specialist* models, allowing them to

achieve further performance boosts. Notably, our foundation models consistently deliver state-of-the-art results on seven benchmarks such as AGORA, UBody, EgoBody, and our proposed SynHand dataset for comprehensive hand evaluation.

Index Terms—Expressive human mesh recovery, foundation model, human pose estimation, 3D human from monocular images, 3D pose, whole-body motion capture.

I. INTRODUCTION

THE innovative realm of 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 uses parametric human models (e.g., SMPL-X [1]) as a powerful representation of the human body, face, and hands. With a flurry of diverse datasets entering the scene in recent years [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], 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 channel their attention towards advancements in architectural designs and 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 first mission in this study is to explore existing data resources extensively, providing key insights crucial for establishing robust, universally applicable models for EHPS. Accordingly, we establish a systematic benchmark for EHPS, utilizing 40 datasets and evaluating their performance across five major benchmarks. Our study reveals 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 paradigm shift towards a more inclusive data selection that offers better in-the-wild performances. Particularly, we propose a new metric, Mean Primary Error (MPE) for both whole-body and hand pose, to gauge model performance across a basket of key benchmarks, in order to provide a holistic measurement of generalization capabilities. Our study underscores the importance of harnessing a multitude of datasets to capitalize on their complementary nature. Moreover, we contribute a new dataset, SynHand, to

Received 9 January 2025; revised 14 September 2025; accepted 1 October 2025. Date of publication 7 October 2025; date of current version 9 January 2026. This work was supported in part by the JSPS KAKENHI under Grant 25H01133 and Grant 25K22112, in part by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 under Grant MOET2EP20221-0012 and Grant MOE-T2EP20223-0002, and in part by cash and in-kind funding from NTU S-Lab and industry partner(s). Recommended for acceptance by J. Denzler. (Wanqi Yin and Zhongang Cai contributed equally to this work.) (Corresponding authors: Lei Yang; Ziwei Liu.)

Wanqi Yin is with The University of Tokyo, Tokyo 113-8654, Japan, and also with SenseTime Research, Singapore 098632.

Zhongang Cai and Hui En Pang are with SenseTime Research, Singapore 098632, and also with S-Lab, Nanyang Technological University, Singapore 637335.

Ruishi Wang, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, and Lei Yang are with SenseTime Research, Singapore 098632 (e-mail: yanglei@sensetime.com).

Ailing Zeng and Lei Zhang are with International Digital Economy Academy (IDEA), Shenzhen 518045, China.

Mingyuan Zhang, Chen Change Loy, and Ziwei Liu are with the S-Lab, Nanyang Technological University, Singapore 637335 (e-mail: ziwei.liu@ntu.edu.sg).

Atsushi Yamashita is with The University of Tokyo, Tokyo 113-8654, Japan. Code is available at: <https://github.com/wqyin/SMPLest-X>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TPAMI.2025.3618174>, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3618174

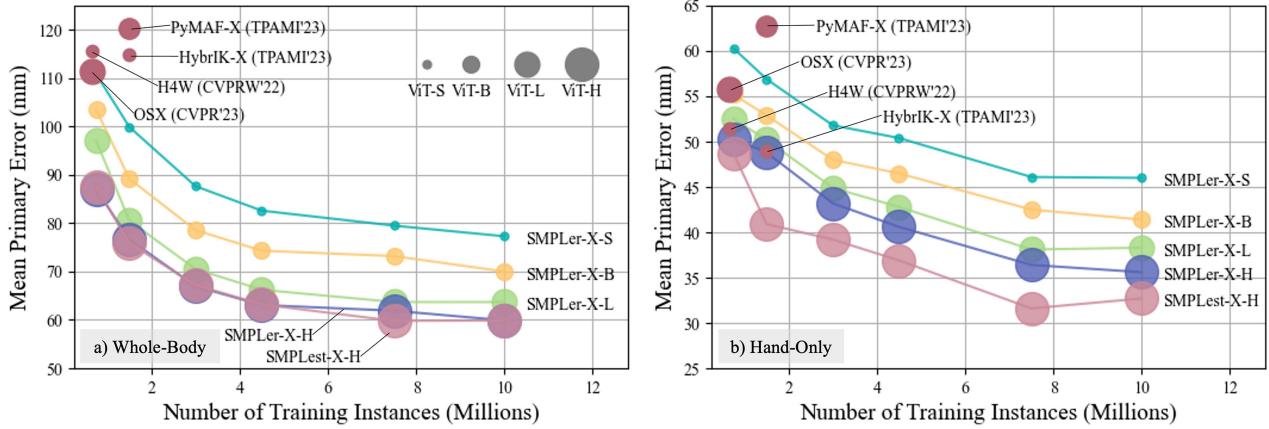


Fig. 1. Scaling up EHPS. (a) Whole-body and (b) hand-only mean primary error (MPE) indicate both data and model scaling are effective in reducing mean errors on primary metrics across key benchmarks for: AGORA [12], UBody [6], EgoBody [5], 3DPW [13] and EHF [14]. OSX [6] and HybrIK-X [15] are SOTA methods. Area of the circle indicates model size, with ViT variants as the reference (top right in the left figure).

provide the community with a long-awaited benchmark for comprehensive hand pose evaluation in a whole-body setting. SynHand features diverse hand poses in close-up human shots, accurately annotated as part of the whole-body SMPL-X labels.

Equipped with the knowledge acquired from the benchmarks, we proceed to exhibit the strength of massively selected data in crafting SMPLer-X and SMPLest-X, a family of *generalist* foundation models that are trained using a diverse range of datasets and achieve exceptionally balanced results across various scenarios. To decouple from algorithmic research work, we intentionally design SMPLer-X and SMPLest-X with a minimalist mindset: they are reduced to very simple architectures. Between them, SMPLer-X opts for a relatively conventional intermediate step for component (hands and face) guiding, while SMPLest-X, as its name suggests, only retains the ultimately essential components for EHPS (an encoder, a decoder, and regression heads). Despite that SMPLer-X and SMPLest-X are designed to facilitate massive data and parameter scaling instead of a rigorous investigation into the algorithmic aspect, we highlight that their simplicity does not undermine their overall performance, and surprisingly, SMPLest-X even demonstrates a special edge on hand estimation in a one-stage manner.

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. Notably, our foundation models demonstrate significant performance boost through both data scaling and model size scaling, reducing the whole-body mean primary errors on five major benchmarks (AGORA [12], UBody [6], EgoBody [5], 3DPW [13], and EHF [14]) from over 110 mm to below 60 mm and the hand primary errors on six major benchmarks from over 62 mm to 31 mm (demonstrated in Fig. 1), and showcases impressive generalization capabilities by effectively transferring to new scenarios, such as DNA-Rendering [11] and ARCTIC [8]. 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

96.2 mm in NMVE, 31.1 mm in hand MVE and achieving SOTA performance on key benchmarks such as EgoBody and UBody.

Our contributions are three-fold. **1)** We systematically study existing EHPS datasets. This includes building the first systematic benchmark on EHPS datasets to provide critical guidance for scaling up the training data toward robust and transferable EHPS, and proposing SynHand as a better whole-body benchmark that focuses on hand evaluation. **2)** We design minimalistic foundation models (SMPLer-X and SMPLest-X) that are tailored for scaling up. Despite their simplicity, they demonstrate capabilities to achieve strong whole-body performance and SMPLest-X further shows promising hand performance in a pure one-stage manner. **3)** We investigate the impact of data and model scaling in EHPS, showing that data scaling yields the largest gains up to 3 M training instances, after which improvements diminish and saturate at 10 M instances. As a result, we build *generalist* foundation model that delivers balanced, strong results across various scenarios and extends successfully to unseen datasets, finetuning further elevates our foundation model into a potent *specialist* across a multitude of key benchmarks.

II. RELATED WORK

Expressive Human Pose and Shape Estimation (EHPS): Due to the erupting expressive avatar applications and the advent of whole-body parametric model (e.g., SMPL-X [14]), capturing the expressive human pose and shape (EHPS), which estimates body, hand, and face together from images, has attracted increasing attention [6], [14], [44], [45], [46], [47], [48], [49], [50]. Optimization-based methods (e.g., SMPLify-X [14]) detect 2D features corresponding to the whole body and fit the SMPL-X model. However, they suffer from slow speed and poor performance. 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 hard. 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 [15], [45], [46], [47], [48], [49], [50], [51]. 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. OSX [6] proposes the first one-stage framework based on ViT-based backbone [52], followed by recent methods such as AiOS [53] and Multi-HMR [54]. This line of work does not require a separate detection stage; it integrates localization and parameter estimation into a single process. As a result, this architecture alleviates the issues in previous multi-stage pipelines, and enhances EHPS performance in scenes prone to imperfect detection. Despite the concise design, they only use confined training datasets for a fair comparison and do not explore the combination of more data to achieve generalizable and precise EHPS.

Foundation Model: The concept of foundation models has had a strong impact across fields [55], [56], [57], [58] and across modalities [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69]. In general vision tasks, SAM [64], [65] introduced promptable segmentation at scale, trained on over one billion masks, and demonstrated strong zero-shot generalization across domains on both images and videos. Beyond segmentation, the DINO family of models [66], [67], [70], [71] has shown that self-supervised training of ViTs with large-scale data and model scaling can produce general representations that transfer effectively to classification, segmentation, and depth estimation tasks. The Depth Anything model family [68], [69] extends the foundation model paradigm to monocular depth estimation, providing scalable pretraining for generalizable depth estimation in real-world scenarios. More recently, VGGT [59], a foundation model in 3D vision trained on a large collection of datasets, achieves state-of-the-art results across multiple 3D tasks and can be further fine-tuned for various downstream applications.

Large-scale Training for Human-centric Vision: Recent efforts use multiple datasets to pretrain a general model for a wide range of downstream human-centric tasks. For example, HumanBench [72] leverages 37 datasets, whereas UniHCP [73] 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. Sapiens [74], trained on 300 million high-resolution human images, achieves strong results in 2D pose estimation, part segmentation, depth estimation, and surface normal prediction, surpassing task-specific baselines. Sárándi et al. [75] 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. [76] analyze 31 datasets for human pose and shape estimation (*i.e.*, SMPL estimation). Nonetheless, hands and face estimation is not included, and only fewer than ten datasets are used concurrently in the most diverse training. HaMeR [77] consolidates 10 datasets, up to 2.7 M training examples for training a transformer model for 3D hand-only reconstruction. NLF [78] is capable of mixing data sources with different annotation formats, including parametric annotations, 3D keypoint annotations, 2D keypoint annotations and DensePose annotations. Despite this, it suffers from inaccurate hand

pose due to the limited sample points on hands. This paper targets the more difficult and essential EHPS task to scale training data and model size that accurately recovers the expressive pose and shape of the human body, hands, and face.

III. BENCHMARKING EHPS DATASETS

A. Preliminaries

SMPL-X: We study expressive human pose and shape estimation via 3D parametric human model SMPL-X [1], 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.

Evaluation Metrics: We use standard metrics for EHPS. PVE (per-vertex error) and MPJPE (mean per-joint position error) measure the mean L2 distances between vertices and regressed joints, respectively. The “PA” prefix indicates Procrustes Alignment is conducted before error computation. AGORA Leaderboard [12] 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 [13] test, and PVE for AGORA-val, UBody, EgoBody, and EHF) to gauge generalizability. Similar to the whole-body counterpart, we propose hand-MPE (the mean of hand PVE) for wrist-aligned hand pose and hand-PA-MPE (the mean of hand PA-PVE) for local pose of hands from multiple datasets (AGORA-val, UBody, EgoBody, EHF, ARCTIC and SynHand) to evaluate the performance of hand pose estimation across key benchmarks. As 3DPW [13] provides the annotation in SMPL format for body-only evaluation, we replace it with ARCTIC [8] for hand-relevant metrics. All errors are reported in millimeters (mm).

B. Overview of Data Sources

In this work, we gather 40 datasets from three major types. 1) motion capture datasets that leverage optical [8], [9], [30], [31], [34], [35], [38] or vision-based [5], [11], [27], [42] 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 or LiDARs [13], [26], [36]. These datasets generally provide high-quality 3D annotations but are less flexible due to physical constraints, especially those built with optical motion capture systems. 2) pseudo-annotated datasets [6], [10], [17], [19], [20], [21], [24], [25], [29], [32], [33], [37], [39], [40], [41], [43], [79] that re-annotate existing image datasets with parametric human annotations. 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 [76]. 3) synthetic datasets [2], [3], [4], [12], [22], [28] 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,

TABLE I

BENCHMARKING EHPS DATASETS. FOR EACH DATASET, WE TRAIN A MODEL ON ITS TRAINING SET AND EVALUATE ITS PERFORMANCE ON THE *TESTING* SETS OF FIVE MAJOR BENCHMARKS: AGORA, UBody, EgoBody (EgoSet), 3DPW, AND EHF. DATASETS ARE THEN RANKED BY MEAN PRIMARY ERROR (MPE), TOP-1 VALUES ARE BOLD, AND THE REST OF TOP-5 ARE UNDERLINED. #INST.: NUMBER OF INSTANCES USED IN TRAINING. ITW: IN-THE-WILD. EFT [16], NEURALANNOT (NEA) [17] AND UP3D [18] PRODUCE PSEUDO LABELS. UNIT: MM.

Dataset	#Inst.	Scene	Real/Synthetic	SMPL	SMPL-X	AGORA [12]		UBody [6]		EgoBody [5]		3DPW [13]		EHF [14]		MPE↓
						PVE↓	★	PVE↓	★	PVE↓	★	MPJPE↓	★	PVE↓	★	
BEDLAM [2]	951.1K	ITW	Syn	-	Yes	164.7	4	132.5	8	109.1	2	98.1	2	81.1	1	117.1
SynHand	462.8K	ITW	Syn	-	Yes	182.3	6	141.2	10	<u>111.0</u>	3	108.35	9	<u>99.1</u>	4	<u>128.4</u>
SynBody [4]	633.5K	ITW	Syn	-	Yes	<u>166.7</u>	5	144.6	12	136.6	8	106.5	7	112.9	8	<u>133.5</u>
InstaVariety [19]	2184.8K	ITW	Real	NeA	-	195	11	<u>125.4</u>	4	140.1	13	<u>100.6</u>	4	110.8	7	<u>134.3</u>
GTA-Human [3]	1802.2K	ITW	Syn	-	Yes	<u>161.9</u>	3	143.7	11	139.2	12	103.4	6	126	15	<u>134.8</u>
MSCOCO [20]	149.8K	ITW	Real	EFT	NeA	191.6	10	<u>107.2</u>	2	139	11	121.2	15	116.3	10	135.0
AGORA [12]	106.7K	ITW	Syn	Yes	Yes	124.8	1	128.4	6	138.4	10	131.1	17	164.6	30	140.6
EgoBody-MVSet [5]	845.9K	Indoor	Real	Yes	Yes	190.9	9	191.4	22	<u>127.0</u>	4	<u>99.2</u>	3	<u>101.8</u>	5	142.1
IDEA400 [21]	1362.9K	Indoor	Real	-	Yes	213.4	23	193.7	23	<u>129.4</u>	5	106.6	8	<u>91.4</u>	2	146.9
WHAC-A-Mole [22]	2500.4K	ITW	Syn	-	Yes	184.6	7	150.7	14	130.8	6	117.7	13	152.8	26	147.3
EgoBody-EgoSet [5]	90.1K	Indoor	Real	Yes	Yes	207.1	19	<u>126.8</u>	5	103.1	1	134.4	23	121.4	13	147.5
RICH [23]	243.4K	ITW	Real	-	Yes	195.6	12	168.1	19	137.9	9	115.5	12	127.5	16	148.9
MP11 [24]	28.9K	ITW	Real	EFT	NeA	202.1	13	<u>123.9</u>	3	155.5	19	131.9	19	140.8	19	150.8
MuCo-3DHP [25]	465.3K	ITW	Real	Yes	-	187.7	8	185.4	21	146.4	16	119.4	14	134.7	18	154.7
EMDB [26]	109.8K	ITW	Real	Yes	-	205.9	17	202.8	25	136.1	7	97.8	1	148.4	23	158.2
PROX [27]	88.5K	Indoor	Real	-	Yes	204.1	15	180.3	20	151.8	17	132.5	22	122.5	14	158.2
UBody [6]	683.3K	ITW	Real	-	Yes	207	18	78.7	1	145.6	15	149.4	29	132.1	17	158.5
SPEC [28]	72.0K	ITW	Syn	Yes	-	<u>161.5</u>	2	146.1	13	154.8	18	139.7	26	197.8	34	160.0
CrowdPose [29]	28.5K	ITW	Real	NeA	-	207.1	20	129.8	7	156.9	20	156.3	31	154.5	27	160.9
MPI-INF-3DHP [30]	939.8K	ITW	Real	NeA	NeA	221.5	26	166.7	18	142.7	14	131.6	18	155.5	28	163.6
HumanSC3D [31]	288.4K	Studio	Real	-	Yes	215.2	24	237.8	28	167.3	21	113	11	107.1	6	168.1
PoseTrack [32]	28.5K	ITW	Real	EFT	-	218.1	25	161	15	180.8	26	150.2	30	149.9	25	172.0
BEHAVE [7]	44.4K	Indoor	Real	Yes	-	208.3	21	205.8	26	175.8	24	132	20	145	21	173.4
CHI3D [9]	252.4K	Studio	Real	-	Yes	203.3	14	264.7	32	175.7	23	122.6	16	121	12	177.5
DAMON [33]	4.4K	ITW	Real	Yes	-	222.2	27	161.2	16	174.8	22	173	32	164.4	29	181.5
Human3.6M [34]	312.2K	Studio	Real	Yes	NeA	226	28	276.1	33	200.6	30	112.3	10	120.8	11	187.2
Hi4D [35]	186.5K	Studio	Real	Yes	-	213.2	22	330.5	37	204.1	32	<u>102.1</u>	5	93.5	3	188.7
DNA-R-HiRes [11]	998.1K	Studio	Real	-	Yes	230	29	278.2	34	179.2	25	134.5	24	149.7	24	194.3
SLOPER4D [36]	33.2K	ITW	Real	-	Yes	205.3	16	255.1	29	190	28	140.3	27	189.2	32	196.0
3DPW [13]	22.7K	ITW	Real	Yes	NeA	234	30	259.3	30	192.6	29	140.6	28	142.9	20	207.2
ARCTIC [8]	1539.1K	Studio	Real	-	Yes	308.5	36	200.7	24	186.4	27	202.5	33	182.5	31	216.1
DNA-R [11]	3992.0K	Studio	Real	-	Yes	274.7	33	341.5	38	214.4	35	138.4	25	115.5	9	216.9
UP3D [18]	7.1K	ITW	Real	UP3D	-	257.5	31	224.1	27	216.6	36	211.5	34	194.8	33	220.9
Talkshow [10]	3326.9K	Indoor	Real	-	Yes	286.4	34	133.2	9	203.6	31	291.3	36	201.9	36	223.3
SignAvatars [37]	4975.1K	Studio	Real	-	Yes	317.4	37	164.6	17	207.2	33	306.82	38	201.47	35	239.5
FIT3D [38]	1779.3K	Studio	Real	-	Yes	329.7	38	404	39	213.8	34	132.1	21	148.1	22	245.5
MTP [39]	3.2K	ITW	Real	Yes	Yes	272.7	32	284.9	35	273.2	37	265.2	35	244.6	37	268.1
OCHuman [40]	2.5K	ITW	Real	EFT	-	307.1	35	263.3	31	279.3	38	293.4	37	281.7	38	285.0
LSPET [41]	2.9K	ITW	Real	EFT	-	365.7	39	292.6	36	340.1	40	339.8	40	316.3	39	330.9
MoYo [42]	1668.4K	Studio	Real	-	Yes	481.8	40	436.4	40	301.7	39	335	39	421.1	40	395.2
SSP3D [43]	311.0K	ITW	Real	Yes	-	549.8	41	522.4	41	548.1	41	439	41	539.5	41	519.8

the synthetic-real gap is not fully addressed. Key attributes of the datasets are included in Table I.

C. Whole-Body Evaluation Benchmarks

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 [12], UBody [6], EgoBody [5], 3DPW [13], and EHF [14]) representing different aspects are selected as the evaluation datasets for whole-body pose evaluation. We briefly introduce the evaluation 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 mainly present 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 transform the estimation in SMPL-X to SMPL topology following [14] for evaluation. **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. 2, we randomly downsample all datasets to equal length (1 K examples) and employ UMAP [80] to visualize several key aspects. We use pretrained ViT-L from HumanBench [72] and OSX [6] 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

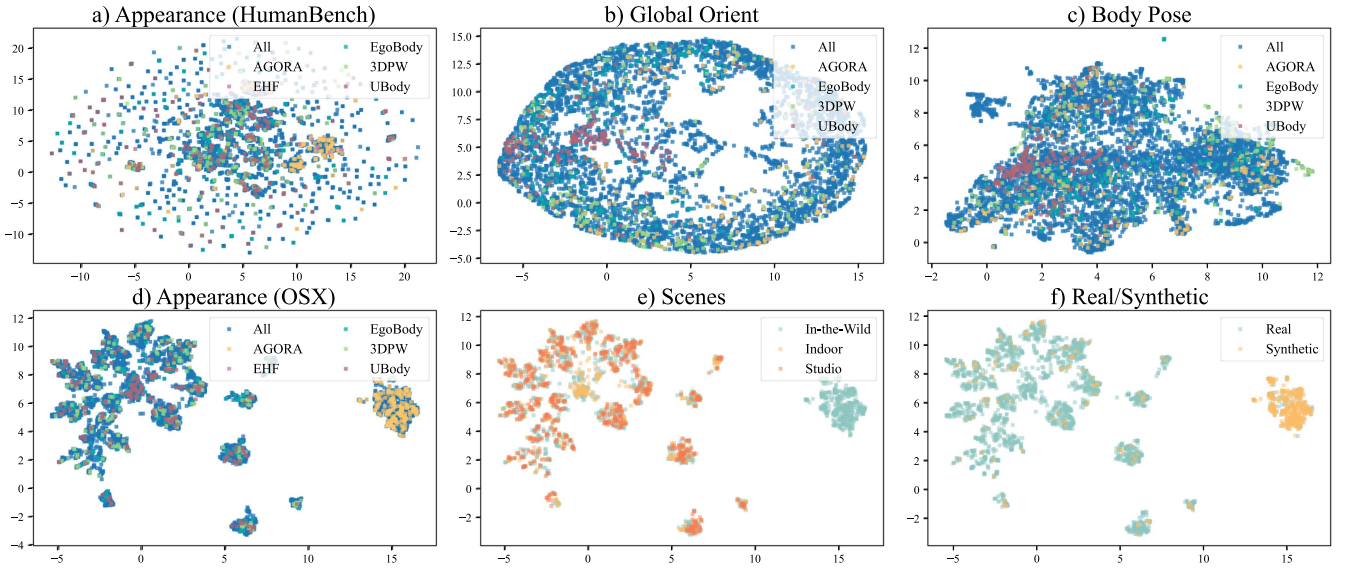


Fig. 2. Dataset attribute distributions. (a) and (d) are image feature extracted by HumanBench [72] and OSX [6] pretrained ViT-L backbone. (b) Global orientation (represented by rotation matrix) distribution. (c) Body pose (represented by 3D skeleton joints) distribution. Both (e) scenes and (f) Real/eot are drawn on the same distribution as (d). All: all datasets. UMAP [80] dimension reduction is used in all visualization with the x and y-axis as the dimensions of the embedded space.

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 [3], [76], [81] 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 mean parameters except for the body pose and hand poses. More evaluation is included in the Supplementary Material. 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.

D. 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 model (SMPLer-X with ViT-S as the backbone, trained for 5 epochs. See Section IV-A for model architecture) with the training set of that dataset and evaluate the model on the *testing* sets of five evaluation datasets: AGORA, UBody, EgoBody, 3DPW, and EHF.

In Table I, we report the primary metrics (Section III-A and ranking of each of the 40 datasets. 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 [10] for gesture generation, DNA-Rendering [11], for human NeRF

reconstruction, MoYo [42] for physical plausibility in complex human motions, and SignAvatars [37] for sign language motion studies), 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, but 6th on UBody, 10th on EgoBody, 17th on 3DPW, and 30th 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.

E. Hand Evaluation Benchmarks

To provide a comprehensive evaluation of the hand performance in the whole-body setting, two additional datasets are selected for hand pose evaluation. First, **ARCTIC** is a lab-based hand-object interaction dataset with accurate SMPL-X annotations for both body pose and hand pose under various scenarios. There are 10 subjects interacting with 11 objects, with 210 K frames of videos captured in total. Second, **SynHand** is the synthetic dataset introduced in this work (elaborated below), which features close-up human shots, diverse body shapes with garments, and a wide range of expressive hand poses.

We introduce **SynHand**, a synthetic dataset specifically designed for EHPS estimation, with a particular emphasis on overcoming the challenges of hand pose estimation with the following motivations. 1) Hand poses in a relaxed state are frequently observed in many whole-body datasets that do not prioritize hands. These datasets often lack specific design or selection criteria for hand poses during the data collection

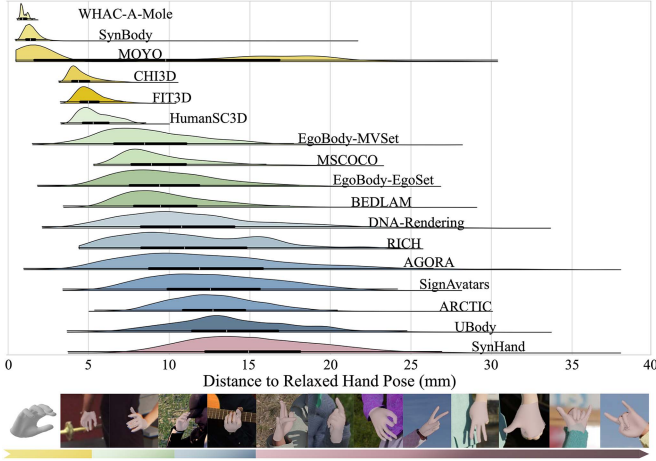


Fig. 3. Analysis of hand poses in whole-body datasets. The distribution of the distance to the relaxed hand pose of each dataset (top) is shown along with the illustration of pose complexity at various distances (bottom). The hand pose with a lower distance is more similar to the relaxed pose.

process. 2) The images in mainstream whole-body datasets, such as those captured or rendered with large fields of view (FOV) [5], [12], often include multiple humans within a single frame. However, this broad FOV comes at the cost of resolution, resulting in low-resolution or barely visible hands in the images. 3) Unlike studio-captured datasets where markers can facilitate in annotations [8], obtaining accurate hand pose annotations remains challenging for in-the-wild datasets.

SynHand leverages the layered human model SMPL-XL [4] and the advanced synthetic rendering toolbox XRFeitoria [83] to generate photorealistic images of virtual humans with diverse body shapes, clothing, and accessories. To capture hands with clarity and visibility, camera placements are strategically optimized for close-up shots, ensuring detailed visual representations. Moreover, the hand poses are extracted from InterHand [84], AGORA [12] and GRAB [85] following BEDLAM [2], while the body poses are sourced from AMASS [86]. For body poses, we carefully select 100 motions from AMASS, in which hand movements can be effectively addressed, such as motions where the wrists are brought forward with fewer occlusions. These motions were chosen to ensure that, with appropriate camera placement, the hand poses can be emphasized. 80% of the body poses are sampled from this subset while 20% are randomly sampled from the entire AMASS dataset. For hand poses, left and right hands are randomly sampled from the three datasets independently. We adopt the official train and test split provided for InterHand datasets. The combined whole-body poses are subsequently used to animate the SMPL-XL model, forming the motion sequences for our dataset. The remaining aspects of data generation, such as scene setup and rendering configurations, follow the approach used in the PDHuman dataset [87]. Fig. 4 shows examples from the SynHand dataset with complicated hand poses and accurate annotations. We also provide the benchmark on SynHand as shown in Table II.

Fig. 3 explains the reason why ARCTIC and SynHand are selected as key benchmarks for hand pose estimation in EHPS.



Fig. 4. Visualization of SynHand dataset with complex hand poses and accurate annotations.

TABLE II
BENCHMARKING EHPS METHODS ON SYNHAND. UNIT: MM. * DENOTES THAT THE METHOD USES INTEGRATED DETECTION RESULTS, F1=98.0

Method	All		Hands		Face	
	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓
Hand4Whole [51]	64.9	116.3	11.7	50.8	3.8	44.2
PyMAF-X [82]	63.5	112.6	11.8	49.2	3.6	44.9
OSX [6]	60.6	107.0	16.7	55.6	4.0	41.6
HybriK-X [15]	53.6	96.5	16.7	50.4	6.0	43.6
AiOS [53]	52.2	94.1	11.8	43.9	3.3	33.5
Multi-HMR* [54]	38.2	70.3	11.6	35.2	3.3	29.7
SMPLeR-X-H32	37.7	64.7	14.0	42.7	3.1	28.5
SMPLeR-X-H40	29.2	47.6	10.3	32.0	2.9	22.7
SMPLeS-X-H32	37.3	63.9	12.4	38.6	3.1	28.1
SMPLeS-X-H40	21.7	34.4	9.8	23.7	2.1	15.4

Fig. 3 presents an analysis conducted across 17 representative whole-body datasets with hand pose annotations. It illustrates the distribution of hand poses as distance to the relaxed hand (when hand pose parameters are all zero). Note that the relaxed hands have naturally curled fingers. Specifically, the average per-vertex position distance from the relaxed hand pose is calculated for each hand pose within each dataset. Higher distance values indicate greater deviations from the relaxed hand pose, suggesting more complicated and potentially more meaningful hand poses. The 17 datasets are thus sorted according to their median distance to the relaxed hand pose. Notably, SynHand exhibits the largest median value away from the relaxed hand, and has a wide range of hand pose distribution, spanning from less than 5 to greater than 35. ARCTIC is also ranked highly on this list.

IV. FOUNDATION MODELS

A. Architecture Design

We highlight that training foundation models are very expensive. Hence, it is infeasible to conduct extensive architectural searches. In this light, we introduce our minimalistic foundation model frameworks, SMPLeR-X and SMPLeS-X that only retain the 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.

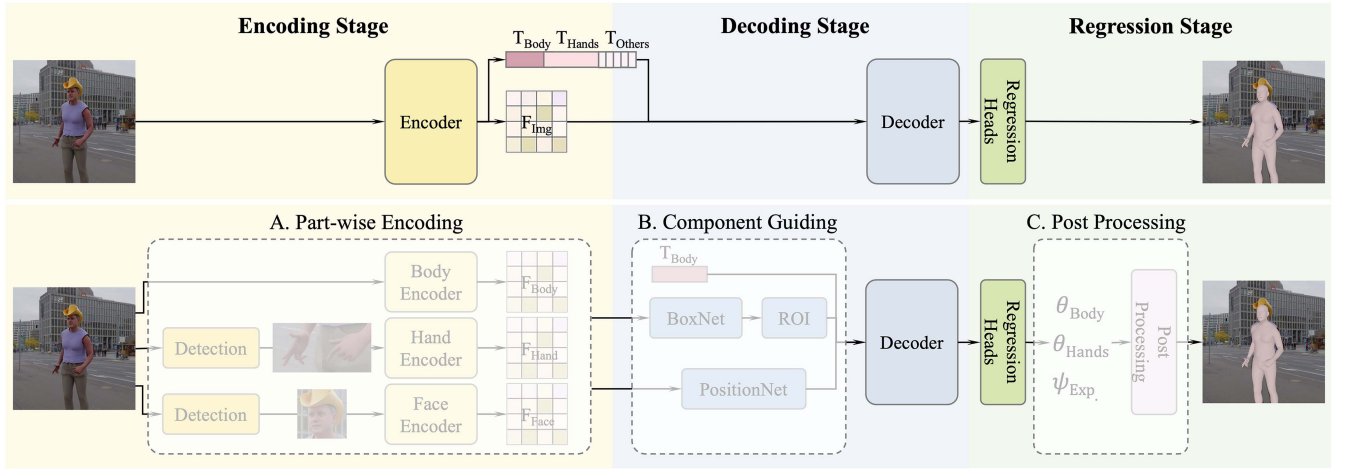


Fig. 5. Architecture of SMPLest-X. Compared with other frameworks with algorithmic modules in various stages (bottom), SMPLest-X (top) has a minimalistic framework design in all three stages. Noted that SMPLer-X consists B. Component Guiding module in the decoder stage.

As illustrated in Fig. 5, the mainstream frameworks can be summarized into three stages: encoding, decoding, and regression stage. Innovative model designs are typically introduced at one or more of these stages. Our minimalistic framework simplifies each stage, ensuring that SMPLer-X and SMPLest-X do not require hand or face decoders following part-wise encoding [46], [51], [82] in the encoding stage. It avoids cross-part feature interaction modules [45], [48], projections of coarse SMPL-X estimations [50] in the decoding stage, and post-processings [15] in the regression stage.

In the decoding stage, SMPLer-X retains only the component guiding module, which is intuitive for by-part estimation. SMPLer-X consists of three parts as shown in Fig. 5: a backbone that extracts image features, which we employ Vision Transformer [52] for its scalability in the encoding stage; the decoding stage consists of a component guiding module that predicts bounding boxes and crops regions of interest from the feature map for hands and face; Multiple regression heads in the regression stage that estimate parameters for each part (e.g. body orientation, body pose, hand pose and expressions). The component guiding module in SMPLer-X functions analogously to a hand and face bounding box detector in feature space. Hand joints regressed from cropped features are used to guide information extraction. However, this leads to significant information loss during cropping from the original encoded features and results in error accumulation during the decoding stage.

To further simplify the design, SMPLest-X eliminates the component guiding module and replaces it with a transformer decoder that has no explicit by-part modeling. SMPLest-X, as its name suggests, is a minimalistic framework that further simplifies SMPLer-X. By replacing the component guiding module with a transformer decoder, we observe a significant performance boost in hand pose estimation.

B. Implementation Details

As shown in Fig. 5, SMPLer-X and SMPLest-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 (256, 192). 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} . For SMPLer-X, the T_{task} is specifically designed for body pose and has a size of $(25, D_{feat})$, where D_{feat} is the depth of the feature. SMPLest-X extends the total size of task tokens to $(80, D_{feat})$ with additional T_{hands} , T_{face} , $T_{orientations}$ and $T_{translations}$ for hand poses, face expressions, body and hands orientations and translations. Note that the hand pose token T_{hands} has the dimension of $(30, D_{feat})$ in total, as the hand pose is considered as complex as the body pose. The tokens are processed with *backbone*. Thanks to the scalability of ViT [52], we are able to experiment with various model sizes. In the decoding stage, *neck* of SMPLer-X processes the image feature F_{img} from the encoder and predicts face and hand bounding boxes. The predicted bounding boxes are used in the ROI (regions of interest) module to crop features from F_{img} , which is re-organized and undergoes transposed convolution (deconv). The component guiding module consists of a positional network to predict 3D keypoints of body and hands. F_{img} with the keypoints guidance will be fed into regression heads along with respective task tokens. Whereas SMPLest-X concatenates F_{img} with T_{task} and feeds it into a 6-layered decoder instead of the component guiding module. The decoded T_{task} is then fed into hand and face heads to predict parameters. In the regression stage of SMPLest-X, we include an additional hand orientation head alongside the heads for body pose, hand pose, and expression. The predicted hand orientation is to formulate a parameter-wise consistency loss with the wrist pose of the body, enhancing accuracy in the distal joints of the upper body and hand positioning.

V. SCALING UP EHPS

A. Generalist Foundation Models

1) *Data Scaling*: The SOTA methods [6], [15] usually train with only a few (e.g., MSCOCO, MPII, and Human3.6 M) datasets, whereas we investigate training with many more

TABLE III

EVALUATE FOUNDATION MODELS ON WHOLE-BODY BENCHMARKS. 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. MPE: MEAN PRIMARY ERROR. THE LOWER, THE BETTER FOR ALL THE METRICS. THE BEST VALUES ARE BOLD AND THE SECOND BEST ARE UNDERLINED. AGORA USES THE VALIDATION SET, AND EGObODY USES THE EGOSet. 40+ DENOTES THE MODEL TRAINED WITH 10 MILLION INSTANCES FROM 40 DATASETS. UNIT: MM.

#Datasets	#Inst.	Model	#Param.	AGORA [12]	EgoBody [5]	UBody [6]	3DPW [13]	EHF [14]	MPE↓
5	0.75M	SMPLer-X-S5	32M	119.0	114.2	110.1	110.2	100.5	110.8
10	1.5M	SMPLer-X-S10	32M	116.0	88.6	107.7	97.4	89.9	99.9
20	3.0M	SMPLer-X-S20	32M	109.2	84.3	70.7	87.5	86.6	87.7
32	4.5M	SMPLer-X-S32	32M	105.2	82.5	68.1	83.2	74.1	82.6
40	7.5M	SMPLer-X-S40	32M	97.9	78.8	71.4	80.2	69.1	79.5
40	10.0M	SMPLer-X-S40+	32M	94.0	78.0	66.9	80.3	67.4	77.3
5	0.75M	SMPLer-X-B5	103M	102.7	108.1	105.8	104.8	96.1	103.5
10	1.5M	SMPLer-X-B10	103M	97.8	76.4	107.3	89.9	74.7	89.2
20	3.0M	SMPLer-X-B20	103M	95.6	75.5	65.3	83.5	73.0	78.6
32	4.5M	SMPLer-X-B32	103M	88.0	72.7	63.3	80.3	67.3	74.3
40	7.5M	SMPLer-X-B40	103M	86.2	70.1	64.4	77.2	68.3	73.2
40	10.0M	SMPLer-X-B40+	103M	81.9	68.2	60.4	74.9	64.6	70.0
5	0.75M	SMPLer-X-L5	327M	88.3	98.7	110.8	97.8	89.5	97.0
10	1.5M	SMPLer-X-L10	327M	82.6	69.7	104.0	82.5	64.0	80.6
20	3.0M	SMPLer-X-L20	327M	80.7	66.6	61.5	78.3	65.4	70.5
32	4.5M	SMPLer-X-L32	327M	74.2	62.2	57.3	75.2	62.4	66.2
40	7.5M	SMPLer-X-L40	327M	69.0	60.7	58.6	71.5	58.6	63.7
40	10.0M	SMPLer-X-L40+	327M	69.2	59.1	54.7	71.1	64.6	63.7
5	0.75M	SMPLer-X-H5	662M	89.0	87.4	102.1	88.3	68.3	87.0
10	1.5M	SMPLer-X-H10	662M	81.4	65.7	100.7	78.7	56.6	76.6
20	3.0M	SMPLer-X-H20	662M	77.5	63.5	59.9	74.4	59.4	67.0
32	4.5M	SMPLer-X-H32	662M	69.5	<u>59.5</u>	54.5	75.0	56.9	63.1
40	7.5M	SMPLer-X-H40	662M	68.2	<u>59.9</u>	56.9	<u>70.9</u>	53.8	61.9
40	10.0M	SMPLer-X-H40+	662M	65.8	57.0	53.7	70.5	52.3	<u>59.9</u>
5	0.75M	SMPLest-X-H5	687M	96.3	92.3	95.2	82.0	71.6	87.5
10	1.5M	SMPLest-X-H10	687M	71.4	68.7	100.8	78.2	60.3	75.9
20	3.0M	SMPLest-X-H20	687M	76.4	63.8	59.8	70.5	65.4	67.2
32	4.5M	SMPLest-X-H32	687M	70.6	59.8	56.5	75.8	53.0	63.2
40	7.5M	SMPLest-X-H40	687M	56.1	61.5	51.1	76.0	49.7	58.9
40	10.0M	SMPLest-X-H40+	687M	<u>59.5</u>	62.0	<u>52.8</u>	74.8	<u>50.5</u>	<u>59.9</u>

datasets. However, we highlight that the dataset benchmark in Table I 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, which ranks individual datasets on the *training* set of the major EHPS benchmarks. We experiment with five dataset counts: 5, 10, 20, 32, and 40 as the training set, corresponding to six different total instance lengths: 0.75 M, 1.5 M, 3.0 M, 4.5 M, 7.5 M, and 10 M (denoted as 40+). For scaling up to 32 datasets, we prioritize higher-ranked datasets in the aforementioned benchmark in Supplementary Material. For scaling up further to 40 datasets, we incorporate additional open-sourced datasets and explore the impact of further increasing instance counts within the same data domain. 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 standardize all datasets into the HumanData format [88].

2) *Model Scaling*: We also study four ViT backbones of different sizes (ViT-Small, Base, Large and Huge). The training is conducted on 16 V100 GPUs, with a total batch size of 512 (256 for ViT-Huge) for 10 epochs (5 epochs for ViT-Huge training

with 7.5 M and 10 M instances from 40 datasets). Specifically, SMPLer-X-L20 and SMPLest-X-L20 take more than 400 GPU hours, SMPLer-X-H32 and SMPLest-X-H32 take more than 700 GPU hours, SMPLer-X-H40+ and SMPLest-X-H40+ take more than 2600 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} .

3) *Evaluating the Generalists*: In Tables III, IV, and V, we show experimental results of data scaling and model scaling with proposed minimalistic framework SMPLer-X and SMPLest-X. It is observed that: 1) More training dataset leads to better performance in terms of MPE, hand-MPE and hand-PA-MPE. The model performance improves gradually as the number of training datasets increases, while with the same training data domain, improving the number of instances does not guarantee the improvement of performance. 2) A larger foundation model performs better at any given amount of data. However, the marginal benefits of scaling up decrease beyond a sufficiently large model size. Specifically, ViT-H has more than twice the parameters than ViT-L, but the performance gain is not prominent. 3) As shown in Fig. 1, Tables III, IV, and V, when the training data is limited (e.g., fewer than 3 M instances), data

TABLE IV

EVALUATE FOUNDATION MODELS ON HANDS. WE STUDY THE SCALING LAW OF THE AMOUNT OF DATA AND THE MODEL SIZES AND EVALUATE THE HAND POSE ESTIMATION. THE METRICS ARE HAND PA-PVE AND HAND PVE FOR ALL EVALUATION BENCHMARKS. (PA-)MPE: (PA-)MEAN PRIMARY ERROR. BEST VALUES ARE BOLDDED AND SECOND BEST ARE UNDERLINED. AGORA USES THE VALIDATION SET, AND EGOBODY USES THE EGOSET. UNIT: MM.

Model	SynHand		AGORA [12]		EgoBody [5]		UBody [6]		ARCTIC [8]		EHF [14]		Hand	
	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓	PA-PVE↓	PVE↓	PA-MPE↓	MPE↓
SMPLer-X-S5	15.6	59.7	10.2	66.8	10.8	53.3	11.8	59.4	16.7	58.7	16.0	64.0	13.5	60.3
SMPLer-X-S10	16.0	54.4	10.2	65.2	10.0	48.6	11.5	57.4	17.5	56.6	16.0	59.1	13.5	56.9
SMPLer-X-S20	15.4	50.8	10.0	63.3	10.0	47.2	11.1	49.6	18.9	45.2	15.5	54.7	13.5	51.8
SMPLer-X-S32	14.6	49.3	9.8	61.9	10.0	46.0	10.7	47.8	18.9	42.9	14.8	54.6	13.1	50.4
SMPLer-X-S40	12.3	42.3	9.1	55.4	9.5	42.7	9.8	44.8	19.2	39.8	13.5	51.6	12.2	46.1
SMPLer-X-S40+	12.7	44.2	9.0	54.6	9.4	42.7	9.6	43.4	19.3	41.1	13.7	50.2	12.3	46.0
SMPLer-X-B5	14.2	53.5	9.6	59.0	10.6	48.0	11.9	56.9	16.9	55.6	15.4	58.4	13.1	55.2
SMPLer-X-B10	14.9	48.7	9.5	57.8	10.1	45.5	12.0	57.1	17.9	53.4	15.7	55.1	13.3	52.9
SMPLer-X-B20	14.1	45.6	9.3	56.5	9.9	44.6	11.0	46.9	18.9	40.9	15.9	53.7	13.2	48.0
SMPLer-X-B32	13.9	44.5	9.2	54.5	9.9	43.7	10.8	43.9	19.0	40.1	14.5	52.1	12.9	46.5
SMPLer-X-B40	11.3	38.3	8.6	50.6	9.2	39.0	9.1	40.8	19.2	37.9	13.4	48.2	11.8	42.5
SMPLer-X-B40+	11.8	38.2	8.5	48.3	9.3	38.8	8.8	38.7	18.9	38.1	13.3	46.3	11.8	41.4
SMPLer-X-L5	13.9	49.6	9.2	53.0	10.5	45.2	12.5	56.3	17.0	52.8	14.7	57.8	13.0	52.5
SMPLer-X-L10	14.7	46.4	9.1	51.9	10.0	43.1	12.8	56.1	18.1	50.8	15.6	52.9	13.4	50.2
SMPLer-X-L20	14.4	43.7	8.9	51.0	9.9	42.7	10.6	43.3	18.9	39.3	15.0	49.4	13.0	44.9
SMPLer-X-L32	13.5	42.4	8.7	47.8	9.8	41.4	10.2	39.2	18.9	38.8	14.1	47.1	12.5	42.8
SMPLer-X-L40	10.7	34.2	8.0	42.8	8.9	36.5	8.4	35.7	19.2	36.1	12.8	43.1	11.3	38.1
SMPLer-X-L40+	11.1	33.0	8.0	43.2	9.0	36.5	8.2	34.2	19.0	36.6	13.3	46.3	11.4	38.3
SMPLer-X-H5	13.4	47.6	9.1	52.6	10.5	43.5	12.1	53.3	17.4	49.3	14.3	55.6	12.8	50.3
SMPLer-X-H10	14.6	44.7	9.0	51.4	10.0	42.6	12.6	54.8	18.8	49.3	15.6	50.2	13.4	48.8
SMPLer-X-H20	14.8	41.9	8.8	49.5	9.8	41.3	10.3	41.0	18.9	38.3	14.4	47.1	12.8	43.2
SMPLer-X-H32	14.0	42.7	8.5	45.6	9.8	39.6	9.8	36.4	18.8	37.0	14.8	42.2	12.6	40.6
SMPLer-X-H40	<u>10.3</u>	32.0	8.0	41.8	<u>8.9</u>	<u>34.8</u>	<u>8.1</u>	34.7	18.9	35.0	12.1	39.9	11.0	36.4
SMPLer-X-H40+	10.8	31.5	<u>7.9</u>	40.1	8.8	34.7	7.9	<u>33.1</u>	19.1	36.3	11.9	38.1	11.1	35.6
SMPLest-X-H5	15.0	48.7	<u>15.2</u>	57.1	10.7	44.4	13.1	48.9	13.2	39.6	14.8	53.6	13.7	48.7
SMPLest-X-H10	12.6	39.8	8.3	41.4	9.2	35.9	11.4	49.0	13.2	37.6	12.9	42.0	11.3	40.9
SMPLest-X-H20	12.6	41.6	8.9	45.2	9.2	37.0	9.5	37.7	11.9	28.8	13.1	44.8	10.9	39.2
SMPLest-X-H32	12.4	38.6	8.2	43.6	9.0	35.7	9.0	35.3	11.6	26.5	12.1	41.5	10.4	36.9
SMPLest-X-H40	9.8	23.7	7.3	36.9	8.8	35.6	7.9	32.9	10.9	<u>24.1</u>	11.4	36.4	9.3	31.6
SMPLest-X-H40+	10.8	<u>25.1</u>	<u>7.9</u>	<u>38.4</u>	9.0	36.3	7.9	34.0	<u>11.1</u>	23.6	<u>11.7</u>	<u>38.8</u>	<u>9.7</u>	<u>32.7</u>

TABLE V

MEAN PRIMARY ERROR (MPE) OF WHOLE-BODY AND HAND POSE ESTIMATION. WE EVALUATE EHPS METHODS ON MULTIPLE BENCHMARK DATASETS FOR WHOLE-BODY AND HAND POSE ESTIMATION. UNIT: MM.

Method	#Params	MPE↓	Hand-MPE↓	Hand-PA-MPE↓
PyMAF-X [82]	206M	120.3	62.7	11.7
Hand4Whole [51]	78M	115.6	51.4	11.4
HybrIK-X [15]	76M	115.0	49.0	14.0
OSX [6]	312M	111.6	55.8	14.0
AiOS [53]	70M	70.6	41.9	11.4
SMPLer-X-S40+	32M	77.3	46.0	12.3
SMPLer-X-B40+	103M	70.0	41.4	11.8
SMPLer-X-L40+	327M	63.7	38.3	11.4
SMPLer-X-H40+	662M	59.9	35.6	11.1
SMPLest-X-H40	687M	58.9	31.6	9.3

scaling provides the most efficient performance improvements. However, beyond the tipping point of 3 M instances, the benefits of additional data gradually diminish, reaching a saturation point at 10 M instances. In this regime, model scaling becomes more effective, enabling the model to better exploit the available data and capture richer information. 4) The foundation model always performs better than in-domain training on a single training set. For example, SMPLer-X-B20, performs better on the test sets of AGORA-val, UBody, EgoBody, and 3DPW, than models trained specifically on the corresponding training set in Table I. 5) By removing the component guiding module in SMPLer-X and replacing it with a decoder, SMPLest-X has significant improvement in hand pose estimation (15% lower in hand-PA-MPE and 13% lower in hand-MPE) as shown in Table V. In addition, we compare our foundation model with SOTA methods across various scenarios, as shown in Fig. 6 for whole-body qualitative

evaluation and Fig. 7 for diverse hand poses in the circumstance of severe occlusion, object interaction and challenging poses.

Moreover, we show detailed by-part results of body, hands, and face on main benchmarks such as AGORA test set (Table VI), AGORA validation set (Table VII), UBody (Table IX), EgoBody-EgoSet (Table X), EHF (Table VIII), ARCTIC (Table XI) and DNA-Rendering (Table XII). We highlight that the foundation models show strong and balanced performances on all benchmarks.

B. Specialist Models

1) *Finetuning the Specialists:* Training the foundation model with a large number of data is expensive. For example, SMPLer-X-H40+ and SMPLest-X-H40+ takes more than 2600 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, besides the respective training set of the target dataset, we use our dataset selection strategy based on the benchmarking in the Supplementary Material and select four high-ranking datasets on the target train set to finetune the model for five epochs.

2) *Evaluating the Specialists:* We finetune the generalist model with ViT-L backbone to match the current SOTA [6] and the generalist models with ViT-H backbone to explore the limit of SMPLer-X and SMPLest-X specialists. The evaluation of specialists are shown in Tables VI, VII, VIII, IX, X, XI, and XII. Moreover, we evaluate the transferability of our foundation models on EHF (Table VIII), ARCTIC (Table XI) and DNA-Rendering (Table XI). EHF is a test-only benchmark

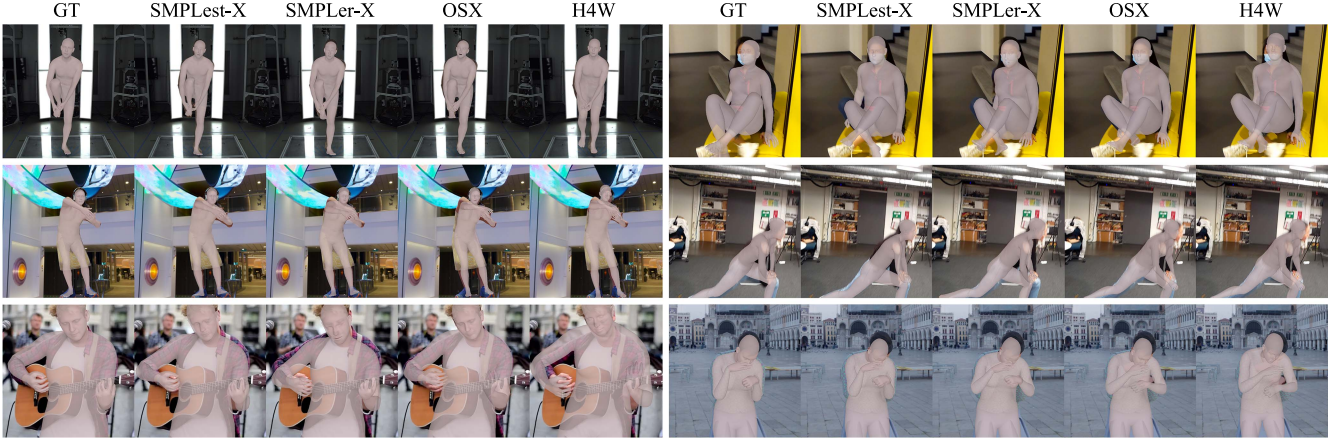


Fig. 6. Visualization. We compare SMPLeSt-X-L40 and SMPLeR-X-L40 with OSX [6] and Hand4Whole (H4W) [51] (trained with the MSCOCO, MPII, and Human3.6M) in various scenarios. Green dotted boxes mark the body parts where our method outperforms others, while red dotted boxes indicate less accurate areas. A zoomed-in version is provided in the Supplementary Material.

TABLE VI

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

Method	NMVE↓ (mm)		NMJE↓ (mm)		MVE↓ (mm)					MPJPE↓ (mm)				
	All	Body	All	Body	All	Body	Face	LHand	RHand	All	Body	Face	LHand	RHand
BEDLAM [2]	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 [51]†	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 [2]†	142.2	102.1	141.0	101.8	103.8	74.5	23.1	<u>31.7</u>	<u>33.2</u>	102.9	74.3	24.7	<u>29.9</u>	<u>31.3</u>
PyMaF-X [50]†	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 [6] *	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 [15]	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
Multi-HMR [54]	102.0	63.4	101.8	64.1	95.9	59.6	27.7	40.2	40.9	95.7	60.3	29.2	38.1	39.0
NLF-L [78]	98.6	<u>62.1</u>	96.6	<u>61.9</u>	92.7	58.4	27.0	37.9	38.1	90.8	58.2	28.5	34.4	34.9
AiOS [53]	97.8	61.3	96.0	60.7	91.9	57.6	24.6	38.7	39.6	90.2	<u>57.1</u>	25.7	36.4	37.3
EMO-X [89]	<u>99.9</u>	<u>63.3</u>	<u>98.4</u>	<u>62.9</u>	<u>91.9</u>	<u>58.2</u>	<u>24.2</u>	<u>38.1</u>	<u>38.9</u>	<u>90.5</u>	<u>57.9</u>	<u>25.1</u>	<u>35.9</u>	<u>36.8</u>
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
SMPLeSt-X-H40†	96.2	63.8	93.9	62.7	86.6	57.4	25.4	30.8	31.5	84.5	56.4	25.7	29.0	29.7

TABLE VII

AGORA VAL SET. † AND * ARE FINETUNED ON THE AGORA TRAINING SET, AND TRAINED ON THE AGORA TRAINING SET ONLY, RESPECTIVELY

Method	PA-PVE↓ (mm)			PVE↓ (mm)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [51]†	73.2	9.7	4.7	183.9	72.8	81.6
PyMAF-X [82]	75.2	12.5	3.8	175.0	119.0	76.5
HybrIK-X [15]	44.6	9.6	4.7	69.0	44.2	35.2
OSX [6]	69.4	11.5	4.8	168.6	70.6	77.2
OSX [6]*	45.0	8.5	3.9	79.6	48.2	37.9
AiOS [53]	43.0	8.5	3.1	63.4	46.1	31.7
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-L20†	39.1	9.3	3.8	62.5	42.3	32.8
SMPLeSt-X-H40	<u>39.2</u>	<u>7.3</u>	<u>2.7</u>	<u>56.1</u>	<u>36.9</u>	<u>26.5</u>
SMPLeSt-X-H40†	37.2	<u>7.8</u>	<u>2.8</u>	<u>57.4</u>	<u>37.7</u>	<u>30.6</u>

TABLE VIII

EHF. AS EHF DOES NOT HAVE A TRAINING SET TO BENCHMARK DATASETS, WE FINETUNE THE MODEL WITH 5 SELECTED DATASETS. AS EHF IS NOT SEEN IN OUR TRAINING AND CAN BE USED TO VALIDATE OUR FOUNDATION MODELS' TRANSFERABILITY.

Method	PA-PVE↓ (mm)			PVE↓ (mm)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [51]	50.3	<u>10.8</u>	5.8	76.8	39.8	26.1
PyMAF-X [82]	50.2	10.2	5.5	64.9	29.7	19.7
HybrIK-X [15]	58.6	16.9	7.3	121.5	53.3	41.4
OSX [6]	48.7	15.9	6.0	70.8	53.7	26.4
AiOS [53]	<u>34.0</u>	12.8	3.8	45.4	44.1	16.9
NLF-L [78]	26.0	20.7	6.3	36.4	43.0	13.9
SMPLeR-X-L20	37.8	15.0	5.1	65.4	49.4	17.4
SMPLeSt-X-H40	34.7	11.4	3.4	49.7	36.4	<u>16.2</u>
SMPLeSt-X-H40†	<u>34.0</u>	11.2	<u>3.7</u>	<u>40.9</u>	<u>33.3</u>	20.2

without training set provided, 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.

C. Training Domains

In Table XIII, we study the impact of using the training split of an evaluation benchmark on its corresponding test split. We have



Fig. 7. Visualization of hands by SMPLeSt-X. SMPLeSt-X demonstrates robust hand pose estimation in whole-body pose and shape estimation tasks across various scenarios, including occlusion (top), object interaction (middle), and challenging hand poses (bottom).

TABLE IX

UBODY. † DENOTES THE METHODS THAT ARE FINETUNED ON THE UBODY TRAINING SET. * DENOTES THE METHODS THAT ARE TRAINED ON UBODY TRAINING SET ONLY.

Method	PA-PVE↓ (mm)			PVE↓ (mm)		
	All	Hands	Face	All	Hands	Face
PIXIE [48]	61.7	12.2	4.2	168.4	55.6	45.2
Hand4Whole [51]	44.8	8.9	2.8	104.1	45.7	27.0
PyMAF-X [82]	78.6	11.2	6.2	146.7	56.0	37.6
HybrIK-X [15]	87.0	14.2	11.2	201.6	57.2	70.7
OSX [6]	42.4	10.8	2.4	92.4	47.7	24.9
OSX [6]†	42.2	8.6	2.0	81.9	41.5	21.2
AiOS [53]	32.5	7.3	2.8	58.6	39.0	19.6
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-L20†	31.9	10.3	2.8	57.4	40.2	21.6
SMPLest-X-H40	27.8	7.9	2.5	51.1	32.9	21.4
SMPLest-X-H40†	26.9	8.4	2.4	50.2	31.8	18.9

TABLE X

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.

Method	PA-PVE↓ (mm)			PVE↓ (mm)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [51]	58.8	9.7	3.7	121.9	50.0	42.5
PyMAF-X [82]	66.3	9.7	3.5	121.8	53.4	45.9
HybrIK-X [15]	52.7	11.2	5.9	111.4	44.6	44.5
OSX [6]	54.6	11.6	3.7	115.7	50.6	41.1
OSX [6]†	45.3	10.0	3.0	82.3	46.8	35.2
AiOS [53]	38.0	9.0	2.9	61.6	40.0	26.7
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-L20†	37.8	9.9	2.9	63.6	42.5	30.8
SMPLest-X-H40	33.6	8.8	2.5	61.5	35.6	28.5
SMPLest-X-H40†	33.5	8.8	2.5	59.3	35.3	28.0

two key observations. First, in-domain training (including the training split of a dataset in training and testing on the test split of the same dataset) is highly effective, as “seeing” the dataset always brings significant performance improvement (R2 vs. R3). Second, having out-of-domain training sets helps if the training

TABLE XI

ARCTIC. † AND * DENOTE THE METHODS THAT ARE FINETUNED ON THE ARCTIC TRAINING SET AND TRAINED ON THE ARCTIC TRAINING SET ONLY, RESPECTIVELY.

Method	PA-PVE↓ (mm)			PVE↓ (mm)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [51]	63.4	18.1	4.0	136.8	54.8	59.2
PyMAF-X [82]	69.4	14.0	3.1	137.9	58.8	53.6
HybrIK-X [15]	63.4	15.4	5.6	147.4	44.5	53.9
OSX [6]	56.9	17.5	3.9	102.6	56.5	44.6
OSX [6]†	33.0	18.8	3.3	58.4	39.4	30.4
AiOS [53]	30.2	19.2	2.1	47.1	38.3	26.1
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-L10†	33.1	19.0	2.7	54.9	40.1	27.3
SMPLest-X-H40	24.2	10.9	3.3	40.3	24.1	28.0
SMPLest-X-H40†	23.2	10.7	3.2	38.1	22.6	27.1

TABLE XII

DNA-RENDERING-HIRES. † AND * ARE FINETUNED ON THE DNA-RENDERING-HIRES TRAINING SET AND TRAINED ON THE DNA-RENDERING-HIRES TRAINING SET ONLY, RESPECTIVELY.

Method	PA-PVE↓ (mm)			PVE↓ (mm)		
	All	Hands	Face	All	Hands	Face
Hand4Whole [51]	62.8	11.0	4.2	111.4	56.4	52.6
PyMAF-X [82]	54.6	11.0	4.3	102.8	55.5	52.6
HybrIK-X [15]	50.7	10.9	5.7	92.0	47.8	48.2
OSX [6]	59.9	10.6	4.3	105.7	55.0	52.5
OSX [6]†	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-L20†	37.9	7.3	3.4	56.5	38.4	34.9
SMPLest-X-H40	33.6	7.4	3.1	56.2	34.6	32.0
SMPLer-X-H40†	32.0	7.1	3.2	51.7	32.8	31.9

data is sufficiently diverse. For example, SMPLer-X-L with four seen datasets benefitted little from having one additional dataset (5 datasets in total with 85.2 mm MPE) but tremendously from having 10 datasets in training (with 72.7 mm MPE). It is worth noting that training with a lot of datasets also benefits out-of-domain (“unseen” benchmarks) performance as errors on EHF (by 20.2 mm PVE reduction), ARCTIC (by 22.8 mm PVE reduction), and DNA-Rendering (by 22.4 mm PVE reduction) decrease with more datasets in the training set and four seen datasets fixed (R3 vs. R5). With the amount of training data limited and fixed, training with out-of-domain data with more varieties brings more benefits than in-domain data (R3 vs. R4). Finally, training on 40 datasets with SMPLest-X-H achieves the best performance with a 58.9 mm MPE, demonstrating it as a robust and effective SMPL-X estimator. However, further increasing the number of instances from 7.5 M to 10.0 M within the same training domain (comprising the 40 datasets) does not promise a performance improvement (R8 vs. R9).

D. Data Sampling Strategies

As for the sampling strategy, we perform 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

TABLE XIII

IMPACT OF IN-DOMAIN TRAINING. 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. THE HIGHLIGHTED YELLOW SHADED NUMBERS DENOTE THAT THE CORRESPONDING TRAIN SPLIT IS USED IN TRAINING. EXCEPT FOR 3DPW USING MPJPE AS THE METRIC, OTHER DATASETS ARE EVALUATED VIA PVE. THE LOWER THE BETTER FOR ALL THE METRICS. TOP-1 VALUES ARE BOLDED, AND THE SECOND BEST VALUES ARE UNDERLINED. EGOBODY: EGOBODY-EGOSET. AGORA: AGORA-VAL. #ROW: ROW NUMBER. #DATA.: NUMBER OF DATASETS. #SEEN: NUMBER OF EVALUATION BENCHMARKS' TRAIN SPLITS USED IN THE TRAINING. #INST.: NUMBER OF TRAINING INSTANCES. UNIT: MM.

#Row	#Data.	#Seen	#Inst.	Model	MPE↓	AGORA [12]	EgoBody [5]	UBody [6]	3DPW [13]	EHF [14]	ARCTIC [8]	DNA-R-HiRes [11]
R1	4	4	0.75M	SMPLer-X-L	83.8	98.1	80.0	65.0	96.2	79.9	101.5	101.3
R2	5	1	0.75M	SMPLer-X-L5	97.0	88.3	98.7	110.8	97.8	89.5	95.1	85.9
R3	5	4	0.75M	SMPLer-X-L	85.2	96.7	81.9	68.1	95.5	83.6	103.6	98.2
R4	10	2	0.75M	SMPLer-X-L10	80.6	82.6	69.7	104.0	82.5	64.0	76.9	76.2
R5	10	4	1.5M	SMPLer-X-L	72.7	84.0	71.6	62.8	81.7	63.4	80.8	75.8
R6	20	5	3M	SMPLer-X-L20	70.5	80.7	66.6	61.5	78.3	65.4	52.2	77.7
R7	32	6	4.5M	SMPLer-X-L32	66.2	74.2	62.2	57.3	75.2	62.3	48.6	54.4
R8	40	6	7.5M	SMPLest-X-H40	58.9	56.1	61.5	51.1	76.0	49.7	40.3	<u>56.2</u>
R9	40	6	10.0M	SMPLest-X-H40+	<u>59.9</u>	<u>59.5</u>	<u>62.0</u>	<u>52.8</u>	74.8	<u>50.5</u>	<u>41.2</u>	56.3

TABLE XIV

DATA SAMPLING STRATEGIES. WE TRAINED SMPLer-X-H32 MODELS WITH DIFFERENT DATA SAMPLING STRATEGIES. UNIT: MM.

Method	Strategy	#Instance	MPE↓
SMPLer-X-H32	balanced	4.5M	63.08
SMPLer-X-H32	weighted	4.5M	62.12
SMPLer-X-H32	concatenated	5.6M	63.32

TABLE XV

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. UNIT: MM.

Method	Dataset	#Instance	MPPE↓
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

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) Concatenated: 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 XIV, while the balanced strategy is more intuitive, easy to implement, and efficient, the weighted strategy may have more potential to improve performance but may require more effort in weight tuning.

E. Training Schemes

As shown in Table XV, we perform the ablation study for the training scheme. We investigate the effect of dataset selection. We select the bottom five and bottom 10 datasets according to the rankings of our individual dataset benchmark on training sets in Supplementary Material 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.

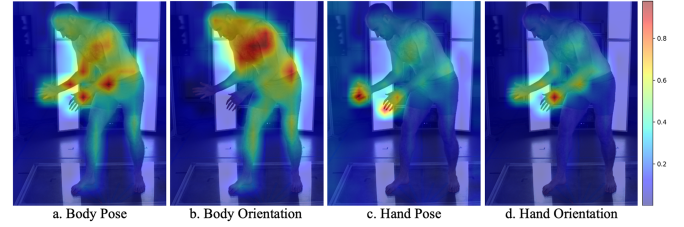


Fig. 8. Attentions of tokens. The extended tokens in SMPLest-X attends to the respective information in the image feature without additional component guidance.

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 five 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.

F. effectiveness of Tokens

In SMPLest-X, the component guiding module from SMPLer-X is replaced with a decoder, and the body pose token is extended to include multiple specific task tokens. As illustrated in Fig. 8, these tokens effectively attend to their respective information in the image feature without requiring additional guidance. For instance, the body pose token activates in the torso region, body orientation tokens focus on the shoulders and pelvis, hand pose tokens concentrate on the fingers of both hands, and hand orientation tokens primarily attend to the upper body, with a particular emphasis on the wrists. The hand pose token has proved effective in capturing local hand pose (evaluated by hand-PA-MPE) while the hand orientation token shows that the upper body is essential for accurate hand orientation (evaluated by hand-MPE).

G. scaling up With CNN Backbones

We choose ViT architecture as our backbone since it has been a popular choice for building large models, and this is the first comprehensive investigation in scaling up both training data and model size of EHPS models. Though the transformer-based

TABLE XVI
SCALING UP WITH CNN BACKBONE. WE STUDY EFFECT OF DATA SCALING ON
THE METHOD WITH CNN BACKBONES. UNIT: MM

Model	Sampling strategy	#Inst.	MPE↓
H4W [51]	-	0.65M	116.59
H4W-32 [51]	balanced	4.5M	98.33
H4W-32 [51]	concatenated	5.6M	96.90

architecture cannot be trivially changed to a CNN-based counterpart (SMPLer-X and SMPLest-X requires task tokens to estimate body pose parameters), we have also trained a Hand4Whole [51], as the CNN-based foundation model (ResNet50 backbone for body and hand, ResNet18 for face with a total of 77 M trainable parameters) as shown in Table XVI to verify that the conclusion still holds true for CNN-based architectures. We highlight that similar to SMPLer-X and SMPLest-X, scaling up the training data for a CNN-based foundation model also shows a substantial performance boost.

VI. DISCUSSIONS

Previous algorithmic studies may have been premature, as the impact of massive data had not been fully explored or ruled out. Our study investigates the non-algorithmic aspects of EHPS through a focus on data- and model-scaling, and systematically apply extensive datasets to their fullest extent on various model sizes. In this section, we discuss the key limitations and take-aways.

By consolidating nearly all available datasets, we observe a clear diminishing return across all model sizes: within the same data domain (40 datasets), increasing the training instances from 7.5 M to 10 M raises the computational cost by 43% (from 23 hours per epoch to 33 hours per epoch), yet results in only marginal performance improvement. Hence, our results suggest that further data scaling has reached a saturation point, indicating the limitations of scaling alone in driving further advancements. With this saturation established, future research on dataset may focus on improving quality through higher-quality annotations and greater diversity, as suggested by our benchmarking analysis. Meanwhile, the field is *better* poised for a necessary shift toward genuine algorithmic advancements upon sufficient pre-training with diverse data.

The minimalistic models, SMPLer-X and SMPLest-X, proposed in this work are parameter-based regression models. However, the improvements from further scaling remain marginal due to the accumulation of errors along the kinematic chain pre-defined in the parametric human model. Future work may focus on designing algorithms to reduce end-effector errors in pose estimation by incorporating inverse kinematics, or on formulating the task as direct mesh recovery for more accurate human shape estimation and improved image–pose alignment.

Moreover, this work adopts the ViTPose [90] family as pretrained backbones, yet more recent human-centric foundation models (e.g., Sapiens [74]) provide pretrained models trained on higher-resolution images and of larger sizes. Such developments

open promising avenues for scaling even stronger foundation models for EHPS.

Another primary limitation of this study is the representativeness of the datasets used. While we employ five comprehensive benchmark datasets to assess whole-body generalization capabilities (with two additional datasets for hand evaluation), these may not fully capture the complexity and diversity of real-world distributions. Nonetheless, to the best of our knowledge, this work constitutes the most systematic generalization study of EHPS to date. We are confident that our observations are broadly applicable to in-the-wild scenarios and provide valuable insights for future research.

VII. CONCLUSION

In this work, we benchmark datasets for EHPS that provide us insights to train and finetune a family of foundation models with minimalistic architectures. Our extensive experiments on data- and model-scaling have systematically explored the impact of leveraging all available datasets to their fullest extent, allowing the foundation models to outperform existing state-of-the-art approaches in accuracy and generalizability.

Our study is thus useful in three ways. First, our systematic study on EHPS datasets, generalization metrics (MPE), and proposed benchmarks serve to gauge the performances of future investigations in building high-performance and robust models for in-the-wild EHPS. Second, SMPLer-X and SMPLest-X are intentionally designed with simplicity and modularity in mind, incorporating highly reusable components and trained on massive datasets, thereby relieving future researchers of the need to replicate this resource-intensive process. Hence, the pretrained models or their components can be a plug-and-play part of any system for EHPS or downstream tasks. Third, our benchmark-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.

REFERENCES

- [1] G. Pavlakos et al., “Expressive body capture: 3D hands, face, and body from a single image,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.
- [2] M. J. Black, P. Patel, J. Tesch, and J. Yang, “Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8726–8737.
- [3] Z. Cai et al., “Playing for 3D human recovery,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10533–10545, Dec. 2024.
- [4] Z. Yang et al., “Synbody: Synthetic dataset with layered human models for 3D human perception and modeling,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 20282–20292.
- [5] S. Zhang et al., “Egobody: Human body shape and motion of interacting people from head-mounted devices,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 180–200.
- [6] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, “One-stage 3D whole-body mesh recovery with component aware transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21159–21168.
- [7] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, “Behave: Dataset and method for tracking human object interactions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15935–15946.

- [8] Z. Fan et al., "Arctic: A dataset for dexterous bimanual hand-object manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12943–12954.
- [9] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu, "Three-dimensional reconstruction of human interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7214–7223.
- [10] H. Yi et al., "Generating holistic 3D human motion from speech," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 469–480.
- [11] W. Cheng et al., "DNA-rendering: A diverse neural actor repository for high-fidelity human-centric rendering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 19982–19993.
- [12] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13468–13478.
- [13] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using imus and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 601–617.
- [14] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.
- [15] J. Li, S. Bian, C. Xu, Z. Chen, L. Yang, and C. Lu, "Hybrik-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2754–2769, 2025.
- [16] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation," in *Proc. Int. Conf. 3D Vis.*, 2022, pp. 42–52.
- [17] G. Moon, H. Choi, and K. M. Lee, "Neuralannot: Neural annotator for 3D human mesh training sets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2299–2307.
- [18] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3D and 2D human representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6050–6059.
- [19] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5614–5623.
- [20] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [21] J. Lin et al., "Motion-X: A large-scale 3D expressive whole-body human motion dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 25268–25280.
- [22] W. Yin et al., "WHAC: World-grounded humans and cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 20–37.
- [23] C.-H. P. Huang et al., "Capturing and inferring dense full-body human-scene contact," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13274–13285.
- [24] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3686–3693.
- [25] D. Mehta et al., "Single-shot multi-person 3D pose estimation from monocular RGB," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 120–130.
- [26] M. Kaufmann et al., "EMDB: The electromagnetic database of global 3D human pose and shape in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 14586–14597.
- [27] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2282–2292.
- [28] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, "SPEC: Seeing people in the wild with an estimated camera," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11035–11045.
- [29] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10863–10872.
- [30] D. Mehta et al., "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 506–516.
- [31] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu, "Learning complex 3D human self-contact," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1343–1351.
- [32] M. Andriluka et al., "Posetrack: A benchmark for human pose estimation and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5167–5176.
- [33] S. Tripathi, A. Chatterjee, J.-C. Passy, H. Yi, D. Tzionas, and M. J. Black, "DECO: Dense estimation of 3D human-scene contact in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 8001–8013.
- [34] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
- [35] Y. Yin, C. Guo, M. Kaufmann, J. Zarate, J. Song, and O. Hilliges, "HI4D: 4D instance segmentation of close human interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17016–17027.
- [36] Y. Dai et al., "Sloper4D: A scene-aware dataset for global 4D human pose estimation in urban environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 682–692.
- [37] Z. Yu, S. Huang, Y. Cheng, and T. Birdal, "SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 1–19.
- [38] M. Fieraru, M. Zanfir, S. C. Pirlea, V. Olaru, and C. Sminchisescu, "AIFit: Automatic 3D human-interpretable feedback models for fitness training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9919–9928.
- [39] L. Müller, A. A. Osman, S. Tang, C.-H. P. Huang, and M. J. Black, "On self-contact and human pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9990–9999.
- [40] S.-H. Zhang et al., "Pose2seg: Detection free human instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 889–898.
- [41] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 1465–1472, 2011.
- [42] S. Tripathi, L. Müller, C.-H. P. Huang, T. Omid, M. J. Black, and D. Tzionas, "3D human pose estimation via intuitive physics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4713–4725.
- [43] A. Sengupta, R. Cipolla, and I. Budvytis, "Synthetic training for accurate 3D human pose and shape estimation in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2020. [Online]. Available: <https://www.bmvc2020-conference.com/assets/papers/0081.pdf>
- [44] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10965–10974.
- [45] G. Pavlakos et al., "Monocular expressive body regression through body-driven attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 20–40.
- [46] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1749–1759.
- [47] Y. Zhou, M. Habermann, I. Habibie, A. Tewari, C. Theobalt, and F. Xu, "Monocular real-time full body capture with inter-part correlations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4811–4822.
- [48] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 792–804.
- [49] Y. Sun, T. Huang, Q. Bao, W. Liu, W. Gao, and Y. Fu, "Learning monocular mesh recovery of multiple body parts via synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 2669–2673.
- [50] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, and Y. Liu, "PyMAF-X: Towards well-aligned full-body model regression from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12287–12303, 2023.
- [51] G. Moon, H. Choi, and K. M. Lee, "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2308–2317.
- [52] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [53] Q. Sun et al., "AiOS: All-in-one-stage expressive human pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1834–1843.
- [54] F. Baradel et al., "Multi-HMR: Multi-person whole-body human mesh recovery in a single shot," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 202–218.
- [55] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, vol. 2022, Art. no. 3876.
- [56] K. Black et al., "π0: A vision-language-action flow model for general robot control," 2024, *arXiv:2410.24164*.

- [57] K. Black et al., “ $\pi 0.5$: A vision-language-action model with open-world generalization,” 2025, *arXiv:2504.16054*.
- [58] B. Zitkovich et al., “RT-2: Vision-language-action models transfer web knowledge to robotic control,” in *Proc. Conf. Robot Learn.*, 2023, pp. 2165–2183.
- [59] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, “VGGT: Visual geometry grounded transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 5294–5306.
- [60] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
- [62] L. Yuan et al., “Florence: A new foundation model for computer vision,” 2021, *arXiv:2111.11432*.
- [63] OpenAI, “GPT-5 system card,” 2025. Accessed: Aug. 7, 2025. [Online]. Available: <https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf>
- [64] A. Kirillov et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
- [65] N. Ravi et al., “SAM 2: Segment anything in images and videos,” in *Proc. 13th Int. Conf. Learn. Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=Ha6RTeWMd0>
- [66] H. Zhang et al., “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=3mRwyG5one>
- [67] M. Oquab et al., “Dinov2: Learning robust visual features without supervision,” *Trans. Mach. Learn. Res.*, 2024. [Online]. Available: <https://openreview.net/forum?id=a68SUt6zFt>
- [68] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth anything: Unleashing the power of large-scale unlabeled data,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 10371–10381.
- [69] L. Yang et al., “Depth anything V2,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 21875–21911.
- [70] O. Siméoni et al., “Dinov 3,” 2025, *arXiv:2508.10104*.
- [71] S. Liu et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 38–55.
- [72] S. Tang et al., “Humanbench: Towards general human-centric perception with projector assisted pretraining,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21970–21982.
- [73] Y. Ci et al., “UniHCP: A unified model for human-centric perceptions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17840–17852.
- [74] R. Khirodkar et al., “Sapiens: Foundation for human vision models,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 206–228.
- [75] I. Sáradi, A. Hermans, and B. Leibe, “Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2956–2966.
- [76] H. E. Pang, Z. Cai, L. Yang, T. Zhang, and Z. Liu, “Benchmarking and analyzing 3D human pose and shape estimation beyond algorithms,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 26034–26051.
- [77] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, “Reconstructing hands in 3D with transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9826–9836.
- [78] I. Sáradi and G. Pons-Moll, “Neural localizer fields for continuous 3D human pose and shape estimation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 140032–140065.
- [79] H. Joo, N. Neverova, and A. Vedaldi, “Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation,” in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 42–52.
- [80] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform manifold approximation and projection,” *J. Open Source Softw.*, vol. 3, no. 29, 2018, Art. no. 861.
- [81] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy, “Delving deep into hybrid annotations for 3D human recovery in the wild,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5340–5348.
- [82] H. Zhang et al., “PyMAF-X: Towards well-aligned full-body model regression from monocular images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12287–12303, Oct. 2023.
- [83] X. Contributors, “Openxrlab synthetic data rendering toolbox,” 2023. [Online]. Available: <https://github.com/openxrlab/xrfeitoria>
- [84] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, “Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 548–564.
- [85] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, “Grab: A dataset of whole-body human grasping of objects,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 581–600.
- [86] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, “AMASS: Archive of motion capture as surface shapes,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5442–5451.
- [87] W. Wang et al., “Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 3925–3935.
- [88] M. Contributors, “Openmmlab 3D human parametric model toolbox and benchmark,” 2021. [Online]. Available: <https://github.com/open-mmlab/mhuman3D>
- [89] H. Jian, J. Zhang, J. Wu, and Z. Tu, “EMO-X: Efficient multi-person pose and shape estimation in one-stage,” 2025, *arXiv:2504.08718*.
- [90] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “Vitpose: Simple vision transformer baselines for human pose estimation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 38571–38584.



Wanqi Yin received the bachelor's degree from Nanyang Technological University, and the master's degree in 2020 from the University of Tokyo, where she is currently working toward the PhD degree advised by Prof. Atsushi Yamashita. She is also an algorithm researcher with SenseTime Research. She has authored or coauthored papers at top-tier conferences including CVPR, NeurIPS, ICCV, and ECCV. Her research interests include computer vision focusing on human motion capture and multi-camera systems.



Zhongang Cai (Member, IEEE) received the bachelor's and PhD degrees from Nanyang Technological University (NTU), advised by Prof. Ziwei Liu and Prof. Chen Change Loy. He is currently a staff research scientist with SenseTime Research. He has authored or coauthored more than 30 papers on top venues, such as NeurIPS, ICLR, CVPR, ICCV, and ECCV. His research focuses on human motion capture and generation. He is actively involved in academic services. He was awarded Lee Kuan Yew Gold Medal as the top student at NTU and Outstanding/Highlighted Reviewer for ICLR'22 and ICCV'21.



Ruisi Wang received the bachelor's degree from the College of Computing and Data Science, Nanyang Technological University. She is currently a researcher with SenseTime Research. She has authored or coauthored several conference papers at CVPR, ECCV. Her research interests include computer vision, deep learning, and human motion capture. She was actively engaged in international competitions, attaining second place in HPC AI Innovation Challenge 2024 and first place in HPC ISC2023 Student Cluster Competition.



Ailing Zeng (Member, IEEE) received the PhD degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong. She was a researcher with IDEA research. She is currently a senior researcher with Tencent AI Lab. She has authored or coauthored more than forty top-tier conference papers at CVPR, ICCV, ECCV, ICLR, ICML, and NeurIPS. Her research interests include computer vision, deep learning, and multi-modal generation, targeting to build multi-modal human-like intelligent agents on scalable Big Data. She was the reviewers, area chairs, or associate editors of many top conferences and journals.



Chen Wei received the bachelor's degree in electric and electronic engineering from Nanyang Technological University. He was an algorithm research intern with S-Lab@NTU, in 2021. He joined SenseTime International Pte. Ltd., in 2022. He has authored or coauthored several papers on top-tier conferences, including ICCV, NeurIPS, and CVPR. His research interests include the generation of synthetic human data from game engines and standardization of human-centric computer vision datasets.



Mingyuan Zhang received the BS degree in computer science and engineering from Beihang University, China. He is currently working toward the PhD degree with MMLab@NTU, advised by Prof. Ziwei Liu. His research interests include computer vision include motion synthesis, 3D pose estimation, and scene understanding. He has authored or coauthored papers on top-tier conferences, including CVPR, ECCV, ICCV, ICLR, and AAAI.



Qingping Sun received the bachelor's degree from Hainan University, and the master's degree from Hunan University. He is currently working toward the PhD degree with the City University of Hong Kong. He is also a research intern with SenseTime Research. His research interests include computer vision and computer graphics.



Lei Zhang (Fellow, IEEE) received the PhD degree in computer science from Tsinghua University, Beijing, China, in 2001. He was the principal researcher and research manager with Microsoft. He is currently the chief scientist of computer vision and robotics with International Digital Economy Academy (IDEA) and adjunct professor with the Hong Kong University of Science and Technology, Guangzhou, China. He has authored or coauthored more than 150 technical papers and holds more than 60 U.S. patents in his research interests which include computer vision and machine learning, with a particular focus on generic visual recognition at a large scale. He was the editorial board member of *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *Multimedia System Journal* and area chair of many top conferences.



Haiyi Mei received the BS degree in automation from the Nanjing University of Science and Technology, and the master's degree in biomedical engineering from Shandong University. He is currently a research fellow with SenseTime. He has authored or coauthored papers in top-tier conferences such as CVPR, ICCV, and NeurIPS. His research interests include text/image-to-video generation and synthetic data for computer vision tasks.



Chen Change Loy (Senior Member, IEEE) received the PhD degree in computer science from the Queen Mary University of London, in 2010. From 2013 to 2018, he was a research assistant professor with the Department of Information Engineering, The Chinese University of Hong Kong. He is currently a professor with the College of Computing and Data Science, Nanyang Technological University. His research interests include computer vision and deep learning with a focus on image/video restoration and enhancement, generative tasks, and representation learning.

He was an associate editor for *Computer Vision and Image Understanding* (CVIU), *International Journal of Computer Vision* (IJCV), and *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI). He was the area chair of top conferences such as ICCV, CVPR, ECCV, ICLR, and NeurIPS.



Yanjun Wang received the bachelor's degree from Shanghai Jiao Tong University, in 2022, under the supervision of Prof. Li Song. He is currently working toward the master's degree in computer science with University of California, Los Angeles. He was a research intern with SenseTime Group Inc., under the supervision of Dr. Lei Yang. He has authored or coauthored multiple papers in relevant fields at international conferences, including CVPR and NeurIPS.



Hui En Pang received the BS degree in biology, and the master's degree in geographic information systems from the National University of Singapore. She is currently working toward the PhD degree in computer science with Nanyang Technological University. She was with Grab, AI Singapore, and SenseTime. She has authored or coauthored in NeurIPS.



Atsushi Yamashita (Senior Member, IEEE) received BE, ME, and PhD degrees in engineering from the University of Tokyo, Japan. He was an associate professor with Shizuoka University, Japan and visiting associate with the California Institute of Technology, USA. He is currently a professor with the University of Tokyo. His research interests include robotics, image processing, and computer vision.



Lei Yang received the BE degree from Tsinghua University, in 2015, and the PhD degree from the Chinese University of Hong Kong, in 2020, advised by Prof. Dahua Lin. He is currently a research director with SenseTime Group Inc. He has authored more than 30 papers on top conferences or journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, SIGGRAPH, IJCV and *IEEE Transactions on Pattern Analysis and Machine Intelligence*.



Ziwei Liu (Member, IEEE) received the PhD degree from the Chinese University of Hong Kong. He was the senior research fellow with the Chinese University of Hong Kong and postdoctoral researcher with the University of California, Berkeley. He is currently an assistant professor with Nanyang Technological University, Singapore. His research interests include computer vision, machine learning and computer graphics. He has authored or coauthored extensively on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Games*, and *Nature - Machine Intelligence*. He was the recipient of Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, and HKSTP Best Paper Award. He was the area chair of ICCV, NeurIPS, and ICLR.