PoseDreamer: Scalable and Photorealistic Human Data Generation Pipeline with Diffusion Models

Lorenza Prospero[1,2]
lorenza@robots.ox.ac.uk
Orest Kupyn[2]
okupyn@robots.ox.ac.uk
Ostap Viniavskyi[3]
viniavskyi@ucu.edu.ua
João F. Henriques[2]
joao@robots.ox.ac.uk
Christian Rupprecht[2]
chrisr@robots.ox.ac.uk

[1] The Podium Institute for Sports Medicine and Technology
University of Oxford
Oxford, UK

[2] Visual Geometry Group
University of Oxford
Oxford, UK

[3] Ukrainian Catholic University
Lviv, Ukraine

arXiv:2603.28763v2 [cs.CV] 3 Sep 2026

Abstract

Acquiring labeled datasets for 3D human mesh estimation is challenging due to depth ambiguities and the difficulty of annotating 3D geometry from monocular images. Existing datasets are either real, with manually annotated 3D geometry and limited scale, or synthetic, rendered from 3D engines that provide precise labels but suffer from limited photorealism, low diversity, and high production costs. In this work, we explore a third path: generated data. We introduce PoseDreamer, a pipeline that leverages diffusion models to generate large-scale synthetic datasets with 3D mesh annotations. Our approach combines controllable image generation with Direct Preference Optimization for control alignment, curriculum-based hard sample mining, and multi-stage quality filtering. Together, these components maintain correspondence between 3D labels and generated images while prioritizing challenging samples to maximize dataset utility. Using PoseDreamer, we generate more than 500,000 high-quality samples, achieving a 76% improvement in image-quality metrics over rendering-based datasets. We find that diffusion-generated and rendered synthetic data are complementary rather than competing: generated data, with its photorealism and in-the-wild variability, benefits performance on in-the-wild scenarios, while rendered data retains an advantage on metrics emphasizing 3D-geometric precision. Combining the two yields the best overall results, and we show this complementarity holds across HMR architectures and training objectives.

1 Introduction

Human recognition tasks in Computer Vision encompass a range of methods for detecting and describing humans in images. These span simpler tasks, such as bounding-box detection, 2D pose estimation, and semantic segmentation, which identify the position and semantics of

© 2026. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.



Figure 1: **PoseDreamer Pipeline:** Our method builds a high-quality synthetic human dataset using aligned controllable diffusion, hard sample mining, and filtering. It produces photorealistic images with precise spatial control and diverse scenarios for robust training. With fixed SMPL-X [46] parameters, varying only the text caption yields diverse, controllable changes in background, clothing, and environment.

humans, to more complex challenges, such as 3D detection, 3D pose estimation, and human avatar reconstruction, which recover the 3D geometry of the human body from images. The success of these methods largely depends on the availability of large, labeled datasets for supervised model training. The Microsoft COCO dataset [33], for example, has significantly advanced 2D human pose estimation. However, creating large labeled datasets is costly and time-consuming, and labeling 3D geometry is more challenging still and prone to error.

A common workaround is to scrape large-scale image collections from the internet and generate pseudo-ground-truth fits. Examples include Instagram images in InstaVariety [23], and YouTube videos in AVA [14]. However, because these images depict real individuals and are not owned by the dataset creators, such datasets raise privacy and consent concerns and can typically be distributed only as links to the original content. This makes them fragile: as hosting platforms remove or relocate content, the dataset degrades over time.

Advanced rendering engines offer one direction, allowing researchers to generate synthetic human datasets with corresponding 3D meshes [5, 10, 45, 66]. However, these approaches face practical limitations that restrict their adoption. First, creating photorealistic renders requires substantial technical expertise in 3D modeling, lighting, and material design. Second, building diverse scenes requires extensive libraries of 3D assets and sophisticated modeling of hair, expressions, and deformations, all computationally intensive. Critically, rendered images often exhibit a "synthetic look" that creates a domain gap when models trained on such data are applied to real-world scenarios, limiting their practical use.

In this work, we explore an emerging alternative: generated data. This approach addresses the concerns raised above. Because the images are synthesized rather than depicting real people, they raise no consent issues, and the dataset can be distributed directly rather than as links to third-party platforms. Generated data also avoids the technical barriers of rendering pipelines and, because large generative models are trained on internet-scale real images, it narrows the domain gap that limits rendered data. This raises the question: has the quality of modern generative models crossed the threshold at which their outputs become useful supervision for downstream 3D tasks? Answering it is the focus of this work. Doing



Figure 2: **Examples from the PoseDreamer Dataset.** High-quality synthetic human samples generated using our pipeline. The examples highlight photorealistic appearance, precise spatial control, and diverse, challenging scenarios that support robust model training.

so, however, requires overcoming a fundamental challenge: ensuring precise correspondence between generated images and their 3D annotations, which existing generative models do not inherently satisfy. We present PoseDreamer, a pipeline for generating synthetic datasets for human mesh recovery that harnesses state-of-the-art diffusion models [3, 31] while ensuring precise 3D-to-2D consistency. We show that diffusion-based generation is a practical alternative to traditional rendering pipelines, offering greater scene diversity and visual realism, and that models trained on PoseDreamer reach performance competitive with those trained on much larger synthetic datasets produced at substantially higher cost. Crucially, we characterize a complementarity between generated and rendered data: generated data improves in-the-wild robustness, while rendered data offers geometric precision; combining them outperforms either alone, a pattern that holds across architectures. Examples of the generated images and annotations are shown in Figure 2. The contributions of this work are:

- **Precise Controllable Human Generation:** We develop an approach for generating photorealistic human images with 3D pose control, introducing an enhanced mesh-to-RGB encoding scheme and employing Direct Preference Optimization to align the control model for improved 3D-to-2D consistency.
- **Curriculum-Based Generation with Model Feedback:** We introduce a two-stage pipeline that incorporates feedback from downstream mesh recovery models, prioritizing challenging samples through hard sample mining and avoiding redundant easy cases.
- **Large-Scale Synthetic Dataset:** We release a synthetic dataset of over 500,000 images,[1] each annotated with a detailed 3D body mesh, providing a diverse resource for training and evaluating human mesh recovery models.
- **Analysis of Generated vs. Rendered Data:** We validate this complementarity across multiple benchmarks, architectures, and training objectives, showing that affordable generation can complement and partially replace costly synthetic data acquisition.

[1] https://prosperolo.github.io/posedreamer

2 Related Work

We divide the literature review into related methods on data generation with diffusion models and an overview of datasets for human mesh recovery.

Data Generation with Diffusion Models. Diffusion models [11, 31] have transformed data generation methods. Recently, [9, 50, 52, 57, 68] improved the performance of classification models by generating synthetic data with latent diffusion models [49]. However, the methods are limited to image classification. Diffumask [65], DatasetDM [64] and S3OD [28] utilize large diffusion models to simultaneously generate synthetic images and annotations for semantic segmentation or depth prediction tasks. Instance Augmentation [27] generates separate objects in an image, providing a framework to augment and anonymize datasets. VGGHeads [29] introduces a first large-scale dataset for human-centric tasks generated with diffusion models. Yet, the data generation process simply predicts ground-truth annotations using state-of-the-art methods, thereby limiting performance. Several recent methods also generate diffusion-based data for human tasks. HumanWild [12] shows that adding generated data (using a SDXL-based ControlNet) to CG-rendered datasets improves performance through greater identity and scene diversity. Diffusion-HPC [63], built on Stable Diffusion, addresses the anatomical inconsistencies often produced by diffusion models through a pose-aware refinement stage. For domain-specific settings, it additionally uses an image-to-image variant that initializes generation from real target-domain images. STAGE [25] builds a similar 3D-pose-conditioned generator but targets the auditing of pose estimators rather than producing training data. Finally, Hayakawa et al. [17] propose a generated dataset for egocentric 3D human pose estimation. In contrast to these approaches, we generate images using mesh-to-RGB conditioning for FLUX, achieve standalone performance competitive with rendered data, and characterize the complementarity between generated and rendered data.

Real-World HMR Datasets. Existing datasets with 3D human annotations face significant limitations despite sophisticated acquisition setups. Human3.6M [20] captures 3.6 million poses using motion capture systems but restricts data to 11 actors performing seventeen scenarios in controlled environments. Similarly, 3DPW [61] provides in-the-wild videos with 3D annotations from video and IMU data, but it is limited to 60 sequences. MPI-INF-3DHP [40] offers 1.3 million frames from 14 synchronized cameras in laboratory settings. While some provide body-shape parameters, their controlled acquisition severely restricts scenario diversity. Pseudo-labeled in-the-wild datasets relax these constraints but inherit the privacy and accessibility concerns discussed above. Several datasets provide full avatar reconstruction of the human body obtained either from complex and expensive capturing systems - CAPE [35] or EHF [46], and have the same pitfalls as the datasets with 3D pose, or use pseudo ground truth data optimized from visual cues, and often suffer from inaccuracies. An example of the latter is NeuralAnnot [42, 43], which labels avatars using ground-truth annotations, such as 2D and 3D joints, available in other datasets.

Synthetic HMR Datasets. Advanced graphics pipelines that render 3D human models are another promising approach for collecting large 3D-annotated datasets. AGORA [45], Surreal [60], and PeopleSansPeople [10] provide hundreds of thousands of fully annotated images. Other approaches like Ultrapose [66] use commercial software but are not publicly available. BEDLAM [5] is a large-scale synthetic dataset produced using physically plausible cloth simulations and artist-generated textures, requiring substantial monetary and computational investment. In contrast, SynBody [67] automates generation with procedural assets, reducing cost at the expense of realism. Concurrently, BEDLAM2 [56] was released,

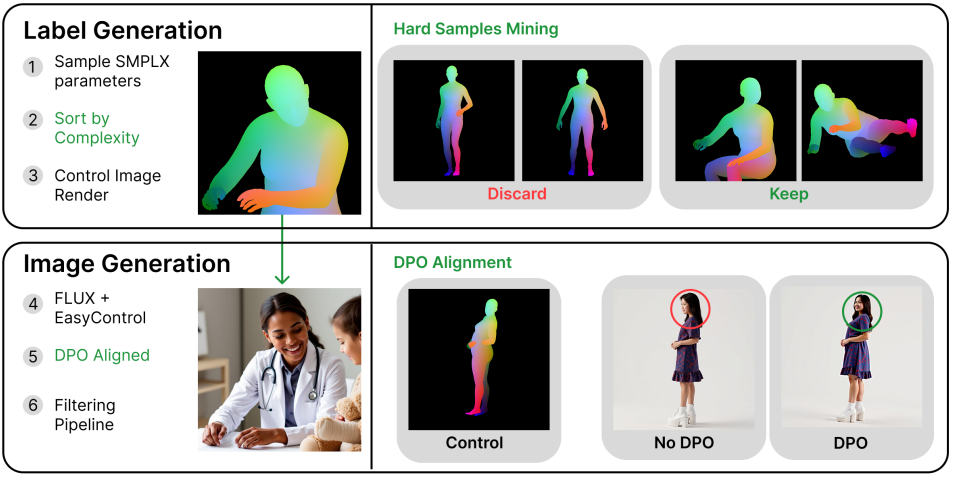


Figure 3: **Data Generation.** Our pipeline begins with SMPL-X [46] parameter and caption generation from multiple sources, followed by curriculum-based hard sample mining to select challenging poses, and concludes with DPO-aligned controllable image generation and comprehensive quality filtering.

targeting video-based estimation with only modest gains for single-frame HMR; we compare against the established BEDLAM. In contrast, our method can scale to arbitrary sample counts and scenarios without manual asset collection.

3 Dataset

Our objective is to construct a large-scale dataset $\mathcal{D}$ defined as: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$, where each pair consists of an RGB image $x_i \in \mathbb{R}^{H\times W\times 3}$ and its corresponding ground truth SMPL-X avatar parameters y_i. Following the established paradigm in human mesh estimation [6], most methods employ a top-down approach: human detection using off-the-shelf detectors, followed by avatar parameter estimation on tightly cropped regions around detected persons. We adopt this framework and focus on single-person mesh recovery.

A straightforward approach would involve generating images $\{x_i\}_{i=1}^{m}$ using latent diffusion models [49], then applying state-of-the-art mesh recovery models to predict corresponding labels $\{y_i\}_{i=1}^{m}$. However, this strategy has a fundamental limitation: current methods lack 3D-2D consistency. When projected onto the original images, the predicted 3D meshes are frequently misaligned with the actual human poses and body shapes. This prevents the use of predictions as reliable ground-truth labels for downstream model training.

To address this limitation, we propose an inverted generation approach. Instead of generating images first, we begin by sampling diverse SMPL-X [46] avatar parameters $\{y_i\}_{i=1}^{m}$ and utilize these as conditioning signals for image generation. Combined with control-model alignment and extensive filtering, this method ensures consistency between the generated images and their associated 3D body parameters. Examples of diverse scenarios generated using the same control are shown in Figure 1, while our complete data generation pipeline is illustrated in Figure 3.

3.1 Generation Pipeline

Label Generation. We generate a set of 3D body model parameters [46] along with detailed captions that describe the corresponding 3D body poses. SMPL-X models human body, hand, and face geometries using a unified parameterization. Specifically, the parameters include pose parameters $\theta \in \mathbb{R}^{55\times3}$; body shape parameters $\beta \in \mathbb{R}^{10}$; and facial expression parameters $\psi \in \mathbb{R}^{10}$. The joint regressor J obtains 3D keypoints from parameters via $R_\theta(J(\beta))$, where R_θ represents the transformation function along the kinematic tree.

A simple approach would randomly sample parameters and condition the generation model on these samples. However, for robust performance and strong generalization, the dataset must contain diverse, realistic scenes with varying interactions, occlusions, and backgrounds, and include a wide range of plausible human 3D poses. To satisfy both requirements, we sample from two complementary sources:

LAION [51]: This large-scale real-world dataset contains over 5B images with diverse scenes and setups, providing sufficient realism and scene diversity but limited pose variety, with predominantly simple standing, sitting, or walking poses. The pipeline begins with YOLO [21] human detection to identify and crop individual persons from the images. We then employ the Gemma3 vision-language model (VLM) [53] to generate detailed captions that describe the visual content, including clothing, environment, and overall scene context. For 3D parameter extraction, we apply a combination of SMPLer-X [6] and Token-HMR [9] models to predict SMPL-X parameters from the cropped human regions. We deliberately combine predictions from both models to mitigate potential bias that could arise from relying on a single downstream model for 3D mesh parameter prediction. This approach helps ensure more robust and reliable ground truth.

AMASS [37]: This database contains over 40 hours of diverse human motion-capture data, providing the challenging, dynamic pose diversity that LAION lacks. The dataset includes complex activities such as dancing, martial arts, sports movements, and everyday interactions that exhibit significant joint articulation and body deformation. For each motion sequence, we sample representative frames and use Gemma [53] to generate detailed captions with specific body-pose descriptions and comprehensive scene descriptions that incorporate diverse environmental elements, lighting conditions, clothing styles, and background settings. The VLM [53] is specifically tasked with describing human pose and generating realistic, detailed scene context, ensuring that the resulting images exhibit wide visual variety while preserving the complex pose structure of the original motion capture data.

Image Generation. To make ground truth sample parameters compatible with state-of-the-art control models [70] and to enforce robust spatial control, we render 3D meshes as RGB images. Common approaches for this conversion include DensePose [15], which maps body surface coordinates to a 2D texture space, and Continuous Surface Embeddings [44], which encode surface correspondences through learned embeddings that maintain consistency across different poses and body shapes.

However, we found that a simple mapping of vertex IDs to RGB colors yields insufficient visual variation across poses, particularly for distinguishing head orientations and body configurations. This limited color variation prevents generative models from incorporating detailed 3D spatial information during conditioning. Building on prior mesh-to-2D encoding approaches [4, 54] and following PNCC notation [71] to address this limitation, we propose an improved color coding scheme normalizing each spatial axis (X, Y, Z) independently and mapping the normalized coordinates directly to RGB channels. Compared to naive mapping,



Figure 4: **DPO Alignment Effectiveness.** Comparison between baseline control model (top row) and DPO-aligned model (bottom row) using identical SMPL-X parameters and captions. The aligned model generates more consistent poses with greater adherence to the ground truth 3D parameters, demonstrating improved control precision and reduced pose deviations.

it provides richer visual cues that better encode spatial relationships and pose variations. Our experimental evaluation, reported in the appendix, shows that models trained with this enhanced mesh-to-RGB mapping achieve better 3D pose generation than baseline color-coding approaches, validating the effectiveness of our spatial encoding strategy.

For the control training, we collect a dataset of 130,000 rendered 3D meshes with their corresponding ground-truth images. To ensure accurate alignment between the rendered meshes and images, we leverage annotations from two existing datasets: DensePoseCOCO [15] and AGORA [45]. Using this dataset, we train a spatial control LoRA [18] following the training procedure of EasyControl [70]. This approach enables the training of a well-aligned control mechanism for image generation. It also provides the flexibility to combine multiple independently trained LoRAs, allowing for more versatile and compositional control.

3.2 Control Model Alignment

While EasyControl [70] provides a mechanism for conditioning image generation on 3D mesh inputs, it represents a weak conditioning method that operates through cross-attention mechanisms without explicit spatial constraints. The model learns to associate control signals with visual patterns during training, but lacks architectural guarantees to enforce precise adherence to the provided 3D pose information, often resulting in generated images that deviate from the intended body configurations.

However, we can effectively detect such misalignments using robust 2D proxy metrics. State-of-the-art 2D joint regressors [38] are significantly more robust than their 3D counterparts, as they are trained on larger datasets and tackle a fundamentally easier task of 2D keypoint localization. Our evaluation setup proceeds as follows: given a generated image and the corresponding ground-truth SMPL-X parameters used for conditioning, we first reproject the SMPL-X model's 3D joints to obtain the ground-truth 2D keypoints. We then apply a 2D joint regressor to the generated image to predict 2D keypoints, and compute the Object Keypoint Similarity (OKS) score between the predicted and reprojected keypoints:

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2\kappa_i^2}\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{1}$$

where d_i is the Euclidean distance between predicted and ground truth keypoints, s is the object scale, κ_i are per-keypoint constants, v_i indicates keypoint visibility, and $\delta(v_i > 0)$ is 1 for visible keypoints and 0 otherwise. This 2D evaluation approach reliably measures pose correspondence while being significantly more stable than direct 3D mesh regression methods, making it ideal for assessing control fidelity.

This capability to reliably measure control alignment enables us to employ Direct Preference Optimization (DPO) [48, 62] to systematically improve the conditioning mechanism. We can enhance control precision and overall generation quality by optimizing the control model to favor generations that exhibit higher pose alignment scores.

We construct preference pairs by ranking generated images according to their OKS scores relative to the ground truth SMPL-X parameters. Images with higher alignment scores are treated as preferred samples (x_0^w), while those with lower scores become less preferred samples (x_0^l), where $x_0^w \succ x_0^l$.

Following the Flow-DPO framework [34], we optimize a control model with:

$$
\begin{aligned}
&-\mathbb{E}\left[\log \operatorname{sigmoid}\left(-\frac{\beta_t}{2}(L_w - L_l)\right)\right] \\
&\text{with } L_w = \|v^w - v_\theta^w\|^2 - \|v^w - v_{\text{ref}}^w\|^2, \quad L_l = \|v^l - v_\theta^l\|^2 - \|v^l - v_{\text{ref}}^l\|^2,
\end{aligned}
\tag{2}
$$

where $\beta_t = \beta(1-t^2)$, and v^* denotes the velocity field for rectified flow models. This objective encourages the control model to generate images that more accurately reflect the input 3D pose constraints, thereby improving control precision and overall generation quality.

We train a separate Low-Rank Adaptation (LoRA) [18] with rank $r = 128$ and $\alpha = 128$. This strategy provides the additional advantage of memory efficiency by eliminating the need to store the reference model separately, as the base model with the adapter disabled naturally serves as p_{ref} during training. Training is conducted with a batch size of 32 for 6,400 iterations using the Adam optimizer with a learning rate of 5×10^{-6}. The complete fine-tuning process requires 2 days on 4 A6000 GPUs.

We validate the effectiveness of alignment on a 1,000-sample held-out test set by measuring control precision using OKS scores between generated images and ground truth 2D keypoints. The baseline control model achieves an OKS of 0.874 ± 0.103, while our DPO-aligned model reaches 0.927 ± 0.052. This represents a substantial **42.1%** reduction in the error rate, demonstrating significant improvements in pose-alignment consistency and reduced variance in generation quality. DPO is thus an efficiency mechanism: it increases the pass rate (by 7% on our held-out subset), avoiding wasted GPU compute on discarded samples. We show examples of the improved OKS alignment in Figure 4.

3.3 Hard Sample Mining

Given an aligned control model and an extensive library of 3D body parameters and captions, we can generate datasets of arbitrary size. However, random sampling often yields redundant samples with similar poses, providing weak training signal. We address this with a curriculum-based two-stage selection that prioritizes challenging examples.

Stage 1: Baseline Generation. We randomly sample N parameter vectors and generate an initial dataset using the aligned control model. We then train an SMPL-X regressor [6] on this dataset and evaluate it on a held-out test set, using OKS between its predictions and ground-truth 2D keypoints. Lower OKS indicates more challenging samples where the regressor underperforms.

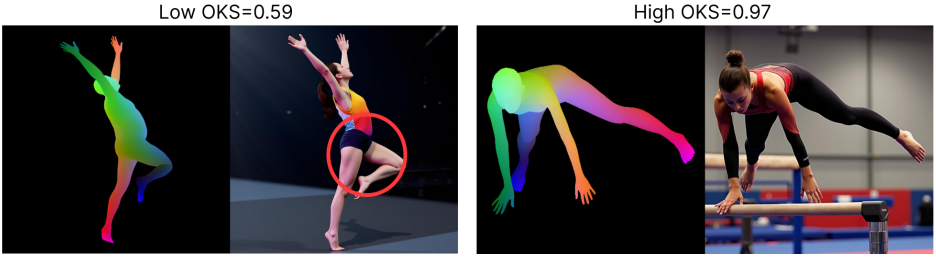


Figure 5: **OKS scores.** Comparison between a low-OKS example (left) and a high-OKS example (right). Each pair shows the conditioning pose alongside the generated image. In the low-OKS case, the generated image has the left and right legs swapped relative to the conditioning pose, causing the detected keypoints to disagree with the projected mesh joints. Only generations with an OKS score above 0.8 pass our filtering stage.

Stage 2: Difficulty-Aware Generation. We train a gradient boosting decision tree regressor [24] to predict OKS scores directly from SMPL-X parameter vectors, letting us rank candidates by expected difficulty without generating images. We then select the top N samples with the lowest predicted OKS and generate the final dataset with the aligned control model. This focuses computation on samples that provide the most learning value while avoiding redundant, easy cases.

This strategy is complementary to pose-source diversity: sources expand the candidate pool, while hard mining selects which candidates to generate given a fixed compute budget.

3.4 Dataset Filtering

Despite our high-quality aligned control model, failure cases inevitably arise. To ensure robust data quality, we implement a comprehensive three-stage filtering pipeline that addresses the primary failure modes we observe in generated data: crowded scenes and severe occlusions, general pose misalignment, and 3D head pose inconsistencies.

Crowded Scenes. We run YOLO [21] on generated images and filter out samples containing more than five detected persons. This threshold removes only highly crowded scenes in which pose estimation quality cannot be reliably ensured, thereby allowing the model to learn partial occlusions and human-to-human interactions.

General Misalignment. We employ the same OKS-based evaluation framework used in control model alignment to filter out samples with poor pose correspondence. Images with OKS scores below a predefined threshold are discarded, some examples are shown in Figure 5. The combination of alignment and filtering is necessary because control model alignment provides key advantages: (a) computational efficiency - aligned models have significantly higher filtering pass rate, reducing wasted computation on poor samples; (b) quality improvement - samples that pass the filtering threshold exhibit better overall alignment when generated from aligned models; and (c) bias mitigation - unaligned models tend to bias toward simple poses that achieve consistency, potentially excluding challenging poses from the final filtered dataset that are essential for robust training.

3D Head Pose Inconsistency. We extract 3D head pose parameters (roll, pitch, yaw) from the ground truth SMPL-X parameters and use VGGHeads [29] to predict head pose from generated images. Samples with mean absolute error (MAE) above a threshold in

any rotation axis are filtered out, ensuring head pose consistency in challenging out-of-distribution poses where general body alignment metrics may be insufficient.

This multi-stage filtering approach ensures that only high-quality, consistent samples contribute to the final training dataset while maintaining diversity in challenging scenarios. The final dataset consists of 500,000 samples filtered from 800,000 initially generated images (500,000 conditioned on LAION poses and 300,000 on AMASS poses). The full generation process required 6,700 GPU hours. This translates to a modest cost compared to traditional rendering pipelines. At July 2026 cloud rates ($0.57/hr A6000, $2.90/hr H100)[2], the full PoseDreamer pipeline (generation + DPO + training) costs ∼$4K. BEDLAM required 111 commissioned simulation-ready garments, cloth simulation, licensed assets, and commercial textures [5]. Public 3D-artist rates ($105–250/garment)[3] already imply $11–27K for garments alone, excluding simulation and asset preparation. Unlike rendering pipelines, PoseDreamer scales primarily with compute rather than specialist labor.

4 Experimental Evaluation

We evaluate PoseDreamer along four dimensions. First, we measure downstream performance by training human mesh recovery models on our generated data and benchmarking their generalization to real-world test sets, both alone and in combination with established synthetic datasets. Second, we compare against existing diffusion-based generation pipelines and show that PoseDreamer can be targeted at specific domains. Third, we assess image quality by comparing PoseDreamer against rendered synthetic datasets on standard photorealism metrics. Finally, we ablate the components of our pipeline to isolate the contribution of each design choice.

4.1 Human Mesh Recovery Benchmarking

We evaluate human mesh recovery along two axes: (i) general benchmarking using the SMPLer-X [6] protocol across diverse academic datasets, and (ii) cross-architecture validation under a different HMR model with a different training objective.

SMPLer-X results. We follow the SMPLer-X [6] benchmarking protocol, training the SMPLer-X-S model on our synthetic data and evaluating its generalization across multiple real-world benchmarks with ground-truth SMPL-X annotations, comparing models trained on single datasets and mixtures. We evaluate on five academic datasets: **AGORA** [45], a synthetic benchmark for rendering engine evaluation; **UBody** [32], covering 15 real-world scenarios with diverse in-the-wild captures; **EgoBody** [69], a large-scale first-person view dataset; **3DPW** [61], providing challenging outdoor sequences; and **EHF** [46], a laboratory dataset with 100 diverse poses.

Table 1 presents results across these benchmarks. Our diffusion-generated dataset achieves strong performance across all evaluations and yields results comparable to BEDLAM [5], which contains roughly twice as many instances and requires substantially greater computational and artistic resources to produce. Notably, our dataset outperforms all real-world datasets in the lower part of the table, which we attribute to the combination of scale, scene diversity, and labels generated natively in the SMPL-X format, whereas real-world

[2] GPU pricing from https://gpuperhour.com, accessed July 2026.

[3] Fiverr 3D artist pricing guide: https://www.fiverr.com/resources/guides/costs/3d-artist, accessed July 2026.

Table 1: **Comparison vs EHPS Datasets.** Top-1 values are bolded, and the rest of Top-2 are underlined. #Inst.: number of instances used in training. ITW: in-the-wild. For each dataset, the model is trained on its training set and evaluated on the *val* set of AGORA and *testing* sets of UBody, EgoBody (EgoSet), 3DPW, and EHF. The upper part of the table shows comparison for models trained on multiple datasets, while the lower part of the table shows comparisons for models trained on a single dataset. Datasets used to train the SMPLer-X-S5 model are marked with ♣.

			AGORA	UBody	EgoBody	3DPW	EHF	
Dataset	#Inst.	Type	PVE↓	PVE↓	PVE↓	MPJPE↓	PVE↓	MPE↓
PoseDreamer + BEDLAM	750K	ITW Mix	156.8	**97.6**	**106.1**	**96.8**	84.3	**108.3**
MSCOCO + BEDLAM	750K	ITW Mix	156.1	101.5	113.4	100.6	**78.7**	110.1
5 datasets (♣)	750K	ITW Mix	**119.0**	110.1	114.2	110.2	100.5	110.8
PoseDreamer	500.0K	ITW Gen	176.0	**98.4**	120.3	102.7	95.2	**118.5**
BEDLAM [8] (our rerun) (♣)	951.1K	ITW Syn	160.5	146.3	**109.5**	**98.7**	**91.2**	121.2
SynBody [57]	633.5K	ITW Syn	166.7	144.6	136.6	106.5	112.9	133.5
InstaVariety [24]	2184.8K	ITW Real	195.0	125.4	140.1	100.6	110.8	134.3
GTA-Human II [2] (♣)	1802.2K	ITW Syn	161.9	143.7	139.2	103.4	126.0	134.8
MSCOCO [33] (♣)	149.8K	ITW Real	191.6	107.2	139.0	121.2	116.3	135.0
EgoBody-MVSet [59]	845.9K	Indoor Real	190.9	191.4	(127.0)	99.2	101.8	142.1
AGORA [35] (♣)	106.7K	ITW Syn	(124.8)	128.4	138.4	131.1	164.6	145.4
Egobody-EgoSet [59]	90.1K	Indoor Real	207.1	126.8	(103.1)	134.4	121.4	147.5
RICH [19]	243.4K	ITW Real	195.6	168.1	137.9	115.5	127.5	148.9
MPII [1]	28.9K	ITW Real	202.1	123.9	155.5	131.9	140.8	150.8
MuCo-3DHP [40]	465.3K	ITW Real	187.7	185.4	146.4	119.4	134.7	154.7
PROX [16]	88.5K	Indoor Real	204.1	180.3	151.8	132.5	122.5	158.2
UBody [34]	683.3K	ITW Real	207.0	(78.7)	145.6	149.4	132.1	158.5
SPEC [23] (♣)	72.0K	ITW Syn	**161.5**	146.1	154.8	139.7	197.8	160.0

datasets typically rely on noisier pseudo-ground-truth fits derived from 2D cues. When combined with BEDLAM, our dataset outperforms all multi-dataset mixtures, including BEDLAM+COCO and the five-dataset ensemble, reflecting complementary strengths: rendered data contributes precise 3D supervision, while generated data contributes photorealistic diversity that rendering pipelines struggle to match. We additionally validate scaling by repeating the evaluation with the larger SMPLer-X-L (ViT-L) backbone, where the same trends hold; full results are provided in the appendix.

Cross-architecture validation. To verify that these findings are not specific to SMPLer-X, we repeat the comparison using HMR2.0 [13], which employs a ViT-H backbone and a keypoint-dominated training objective. We follow the published HMR2.0 architecture and loss configuration, training each model for 250,000 iterations under matched conditions. Each model is trained on the HMR2.0a real-world configuration combined with BEDLAM, PoseDreamer, or both. We use HMR2.0a rather than the full HMR2.0b mix because the additional datasets (InstaVariety, AVA, AI Challenger) suffer from the privacy and accessibility limitations discussed in Section 1. As shown in Table 2, the same complementarity pattern emerges: BEDLAM leads on 3D-geometric and strict-tolerance metrics (3DPW PA-MPJPE and MPJPE), while PoseDreamer and the combined dataset lead on 2D and loose-tolerance metrics (COCO [33], LSPet [22], PoseTrack [1] PCK). The per-metric gaps are smaller than under SMPLer-X, likely reflecting differences in HMR2.0's training configuration. Nonetheless, the appearance-versus-geometry trade-off persists across architectures, supporting it as a property of the data sources themselves rather than a single model artifact.

Table 2: **Cross-Architecture Validation (HMR2.0, ViT-H).** The complementarity pattern holds under a different architecture and training objective: rendered data (BEDLAM) leads on 3D-geometric metrics, while generated data (PoseDreamer) and the combination lead on in-the-wild 2D metrics. Best per column in bold.

Training Data	COCO @.05 ↑	COCO @.10 ↑	LSPet @.05 ↑	LSPet @.10 ↑	PoseTrack @.05 ↑	PoseTrack @.10 ↑	3DPW PA-MPJPE ↓	3DPW MPJPE ↓
Real + BEDLAM	81.82	95.18	46.33	78.11	87.00	97.02	**45.59**	**72.91**
Real + PoseDreamer	82.44	95.27	45.73	78.27	87.94	**97.53**	47.81	75.46
Real + BEDLAM + PoseDreamer	**83.35**	**95.45**	**46.49**	**78.98**	**88.49**	97.44	46.90	74.98

Table 3: PVE (↓) results for models trained on PoseDreamer and BEDLAM on **in-the-wild benchmarks**.

Dataset	UBody	MPII	MSCOCO
BEDLAM [6]	146.3	141.1	163.7
PoseDreamer	**97.6**	**122.3**	**129.4**

Table 4: **PoseDreamer vs. HumanWild** at matched 10K scale. SMPLer-X-S (BEDLAM-pretrained) finetuned on each. IS/FID computed on the 10K subsets.

Dataset	MPII ↓	MSCOCO ↓	UBody ↓	IS ↑	FID ↓
HumanWild [42]	126.21	141.89	110.86	8.10	3.29
PoseDreamer	**124.20**	**134.39**	**101.25**	**11.84**	**1.64**

4.2 Comparison with Diffusion-Based Baselines

HumanWild [42] is, to the best of our knowledge, the only other diffusion-based generic HMR data generation pipeline. It conditions an SDXL-based ControlNet on rendered signals (keypoints, depth, and surface normals) to synthesize images with 3D annotations. As its generated dataset is not publicly released, we perform a matched-scale head-to-head comparison using their publicly released ControlNet: we generate 10,000 samples with each pipeline and finetune SMPLer-X-S (pretrained on BEDLAM) on each. Table 4 shows that PoseDreamer outperforms HumanWild on all three in-the-wild benchmarks and on both image-quality metrics. The largest gap appears on MSCOCO, consistent with the appearance-diversity advantage of PoseDreamer's FLUX-based generation over HumanWild's SDXL base with rendered-normal conditioning. Beyond pose accuracy, the image-quality gap confirms that PoseDreamer produces images closer to real-world distributions than the closest publicly reproducible diffusion baseline.

4.3 Domain-specific Generation

A key advantage of PoseDreamer over traditional datasets is the ability to generate images tailored to a target domain. To evaluate this, we generate 30K images using poses from the MOYO dataset [53] (used exclusively as a pose source, not for retraining), which contains challenging yoga poses underrepresented in other datasets (Figure 6). We train two SMPLer-X-S models: one on 300K randomly sampled AMASS and LAION poses, and one with 30K of these samples replaced with our yoga-pose images. Both are evaluated on UBody, 3DPW, EHF, and the 487 MPII instances tagged "yoga" (Table 6). The yoga-augmented model matches the baseline on general benchmarks while decisively outperforming it on the MPII yoga subset, supported by visual comparisons in Figure 7. Thus, our pipeline generates data tailored to specific domains, improving downstream performance without sacrificing general capability. We additionally compare against Diffusion-HPC [63] on the SMART benchmark [8] of athletic poses; full results are in the appendix.



Figure 6: **Yoga Generations.** Even for the challenging yoga poses from the MOYO [58] dataset, our pipeline generates realistic images with diverse backgrounds.

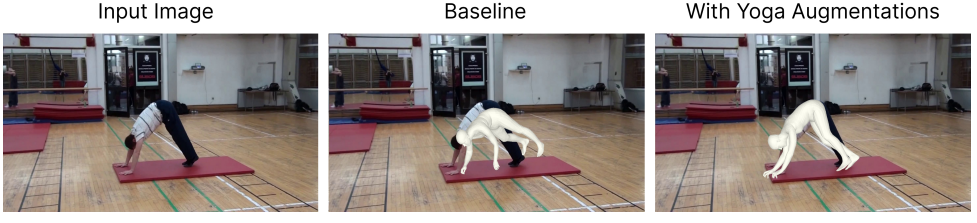


Figure 7: **Yoga-Augmented Training.** Yoga-augmented training improves pose recovery on MPII yoga images.

4.4 Image Quality Evaluation

A critical advantage of our approach is the substantial improvement in visual realism compared to datasets from traditional rendering engines. PoseDreamer significantly outperforms BEDLAM [6], SynBody [67], and AGORA [45] on both image-quality metrics, improving Inception Score by 76% over the best baseline (AGORA, 5.55 $\rightarrow$ 9.78) and reducing FID by 62% over the best baseline (BEDLAM, 4.55 $\rightarrow$ 1.72); full metrics in the appendix. This realism reduces the domain gap that limits rendering-engine approaches. To further validate this, we compare models trained on PoseDreamer and BEDLAM (Table 1) on two additional in-the-wild benchmarks, MPII [1] and MSCOCO [33]. As shown in Table 3, the model trained on PoseDreamer outperforms BEDLAM on these and on UBody, consistent with the appearance-diversity advantage of generated data on in-the-wild scenarios.

4.5 Ablation Study

To evaluate the impact of our pipeline design choices, we compare the downstream model performance on the test set of UBody [32], EgoBody [69], and EHF [46]. We use SMPLer-X [6], trained for 6 epochs on datasets generated with different versions of our pipeline, and report the average PVE metric; the results are in Table 5. Removing any component degrades downstream performance: quality filtering prevents training on misaligned samples, AMASS [37] adds pose complexity, and hard sample mining ensures learning from the most challenging configurations.

Table 5: **Ablations.** We evaluate the impact of our pipeline components based on their effect on downstream model performance.

Dataset	UBody PVE ↓	Ego PVE ↓	EHF PVE ↓
PoseDreamer	**104.4**	**128.0**	**95.1**
w/o Filtering	111.1	130.7	95.4
w/o AMASS [31]	105.3	133.0	103.5
w/o Hard Mining	107.9	131.2	95.4

Table 6: **Domain-Specific Generation.** Replacing 30K of 300K random poses with yoga poses (MOYO) decisively improves MPII Yoga performance with no loss on general benchmarks.

Dataset	UBody PVE ↓	3DPW PVE ↓	EHF PVE ↓	MPII Yoga PVE ↓
270K + 30K yoga	**97.0**	**112.7**	105.3	**171.1**
300K random	97.4	113.3	**104.8**	199.6

5 Limitations and Broader Impact

PoseDreamer offers a complementary alternative to rendering engine datasets with distinct trade-offs. While rendering engines provide exact 3D ground truth, our diffusion-based approach offers greater flexibility, scalability, and photorealism, at the cost of reduced 3D-geometric precision. Despite extensive alignment and filtering, generated samples may contain subtle inconsistencies that rendering engines avoid. As our experiments show, this trade-off is benchmark-dependent: generated data benefits in-the-wild appearance variation, rendered data retains 3D precision, and combining them yields the best overall performance. Improving fine anatomical fidelity for hands and facial expressions through region-specific supervision is a promising direction for future work. Our reliance on foundation generative models may also introduce biases from their training data, which could limit pose diversity for underrepresented demographic groups or activities. Generating realistic human imagery raises ethical concerns about potential misuse for misleading content or unauthorized representations. Because our pipeline conditions on 3D mesh parameters rather than identity-specific features, the generated images do not depict specific real individuals, reducing privacy risk relative to scraping-based pipelines. Our scalable approach broadens access to high-quality human pose datasets, advancing research in motion analysis, healthcare, and assistive technologies while benefiting researchers who lack access to expensive motion capture systems.

6 Conclusion

PoseDreamer demonstrates that diffusion-based synthetic data has crossed the utility threshold for human mesh recovery: models trained on our generated data reach performance competitive with much larger rendered datasets, and the two data sources are complementary, generated data benefiting in-the-wild appearance while rendered data retains 3D precision. Our pipeline combines DPO-based control alignment, PNCC-style mesh-to-RGB conditioning, curriculum-based hard sample mining, and multi-stage OKS filtering to maintain 3D-to-2D correspondence at scale, producing a 500K-image dataset whose utility holds across two HMR architectures and several benchmarks. This positions diffusion-generated data as a viable, cost-efficient alternative that scales with compute rather than commissioned artist labor. As diffusion models continue to improve, we expect this approach to scale further in both quality and diversity while maintaining its cost advantage over rendering pipelines.

Acknowledgments

This work was funded by The Podium Institute for Sports Medicine and Technology, at the University of Oxford. O.K. is supported by a Google unrestricted gift. J.H. acknowledges the generous support of Toyota Motor Europe (TME) and EPSRC (VisualAI, EP/T028572/1). C.R. is supported by ERC StG 101222037-Volute.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018.

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *Transactions on Machine Learning Research*, 2023.

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020.

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

[6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D human recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10533–10545, 2024. doi: 10.1109/TPAMI.2024.3450537.

[8] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021.

[9] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024.

[10] Salehe Erfanian Ebadi, Saurav Dhakad, Sanjay Vishwakarma, Chunpu Wang, You-Cyuan Jhang, Maciek Chociej, Adam Crespi, Alex Thaman, and Sujoy Ganguly. Psp-hdri+: A synthetic dataset generator for pre-training of human-centric computer vision models. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

[12] Yongtao Ge, Wenjia Wang, Yongfan Chen, Fanzhou Wang, Lei Yang, Hao Chen, and Chunhua Shen. Diffusion models are efficient data generators for human mesh recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2025. doi: 10.1109/TPAMI.2025.3621755.

[13] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14737–14748, 2023.

[14] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[15] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[16] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019.

[17] Kyohei Hayakawa, Dong-Hyun Hwang, Chen-Chieh Liao, and Hideki Koike. Diffusion-based synthetic dataset generation for egocentric 3d human pose estimation. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, editors, *Computer Vision – ECCV 2024 Workshops*, pages 236–251, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-91907-7.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

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

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[21] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL https://github.com/ultralytics/ultralytics.

[22] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1465–1472, 2011.

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

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[25] Nikita Kister, István Sárándi, Jiayi Wang, Anna Khoreva, and Gerard Pons-Moll. Are pose estimators ready for the open world? STAGE: A genAI toolkit for auditing 3D human pose estimators. In *International Conference on 3D Vision (3DV)*, March 2026.

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

[27] Orest Kupyn and Christian Rupprecht. Dataset enhancement with instance-level augmentations. In *European Conference on Computer Vision (ECCV)*, pages 384–402. Springer, 2024.

[28] Orest Kupyn, Hirokatsu Kataoka, and Christian Rupprecht. S3OD: Towards generalizable salient object detection with synthetic data. In *International Conference on Learning Representations (ICLR)*, 2026.

[29] Orest Kupyn, Eugene Khvedchenia, and Christian Rupprecht. VGGHeads: 3D multi head alignment with a large-scale synthetic dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2026.

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

[31] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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

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

[34] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 82155–82192, 2025.

[35] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[36] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3D human mesh estimation from virtual markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 534–543, 2023.

[37] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019.

[38] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022.

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

[40] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

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

[42] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural annotator for 3D human mesh training sets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[43] Gyeongsik Moon, Hongsuk Choi, Sanghyuk Chun, Jiyoung Lee, and Sangdoo Yun. Three recipes for better 3D pseudo-gts of 3D human mesh estimation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[44] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33:17258–17270, 2020.

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

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[48] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[50] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8011–8021, 2023.

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[52] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 769–778, 2023.

[53] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

[54] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–110, 2012.

[55] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[56] Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[57] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. StableRep: Synthetic images from text-to-image models make strong visual representation learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[58] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[59] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[60] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

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

[62] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[63] Zhenzhen Weng, Laura Bravo-Sánchez, and Serena Yeung-Levy. Diffusion-HPC: Synthetic data generation for human mesh recovery in challenging domains. In *International Conference on 3D Vision (3DV)*, pages 257–267, 2024. doi: 10.1109/3DV62453.2024.00043.

[64] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. DatasetDM: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023.

[65] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023. doi: 10.1109/ICCV51070.2023.00117.

[66] Haonan Yan, Jiaqi Chen, Xujie Zhang, Shengkai Zhang, Nianhong Jiao, Xiaodan Liang, and Tianxiang Zheng. UltraPose: Synthesizing dense pose with 1 billion points by human-body decoupling 3D model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10891–10900, 2021.

[67] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. SynBody: Synthetic dataset with layered human models for 3D human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20282–20292, 2023.

[68] Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. ImageNet-D: Benchmarking neural network robustness on diffusion synthetic object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21752–21762, 2024.

[69] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022.

[70] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. EasyControl: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19513–19524, 2025. doi: 10.1109/ICCV51701.2025.01814.

[71] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016.

A Additional Analysis

A.1 Scaling

We evaluate PoseDreamer's performance on larger models by training several versions of SMPLer-X-L [6], a ViT-L-based model. The results are shown in Table 7. The model trained on 0.75M samples from two datasets (PoseDreamer and BEDLAM) outperforms

the SMPLer-X-L model trained on a mixture of five datasets across all benchmarks, except EgoBody and AGORA (the latter included in the five-dataset train set).

Table 7: **Scaling.** We train multiple versions of the SMPLer-X-L and SMPLer-X-S models to demonstrate that PoseDreamer contributes to performance improvements when scaling both data and model size.

#Datasets	#Inst.	Datasets	Backbone	AGORA PVE ↓	EgoBody PVE ↓	UBody PVE ↓	3DPW MPJPE ↓	EHF PVE ↓
2	0.75M	Ours + BEDLAM (50-50 split)	ViT-S	156.8	**106.1**	**97.6**	**96.8**	**84.3**
5	0.75M	SMPLer-X-S 5 datasets	ViT-S	119.0	114.2	110.1	110.2	100.5
2	0.75M	Ours + BEDLAM (500K-250K)	ViT-L	138.7	101.1	**93.8**	**88.5**	**73.3**
2	0.75M	Ours + BEDLAM (50-50 split)	ViT-L	138.1	114.2	107.4	100.1	93.9
5	0.75M	SMPLer-X-L 5 datasets	ViT-L	88.3	**98.7**	110.8	97.8	89.5
2	1.5M	Ours + BEDLAM (50-50 split)	ViT-L	131.3	99.6	**94.6**	85.6	67.5
10	1.5M	SMPLer-X-L 10 datasets	ViT-L	82.6	69.7	104.0	**82.5**	**64.0**

Notably, the model trained on an equal split of PoseDreamer and BEDLAM underperforms compared to the model trained on 250K samples from BEDLAM and 500K samples from PoseDreamer. This indicates that PoseDreamer's diversity and photorealism continue to improve performance at scale, while BEDLAM's contribution appears to saturate.

We also compare the SMPLer-X-L model trained on 1.5M samples from two datasets (PoseDreamer and BEDLAM) with a version trained on a mixture of ten datasets. The results are comparable across all benchmarks not included in the ten-dataset training set, that is, every benchmark except AGORA and EgoBody. Moreover, the model trained on 2 datasets outperforms the one trained on 10 datasets on the in-the-wild UBody benchmark.

These results demonstrate that our dataset provides substantial diversity in poses and environments, leading to improved model performance when scaling both model size and the number of training samples.

A.2 Comparison with Diffusion-HPC on SMART

To complement the yoga experiment in the main paper, we compare PoseDreamer against Diffusion-HPC [63] on the SMART dataset [8] of challenging athletic poses. We follow the experimental setup reported in their paper: starting from approximately 100 real images per sport, we generate 3× augmented synthetic images and use them to finetune the downstream regressor, comparing against the numbers reported by Diffusion-HPC. For the SMART experiments, Diffusion-HPC's released setup initializes generation from real target-domain images, so its outputs retain elements of real-image appearance. PoseDreamer, by contrast, generates from scratch conditioned only on pose, with no exposure to the target distribution. We further apply our OKS-based filtering, retaining only generations with an OKS score above 0.8, so the downstream model is trained on fewer but higher-quality images. As shown in Table 8, PoseDreamer matches or exceeds Diffusion-HPC on four of six actions despite this difference in setup. Figure 8 shows examples of generations from SMART poses, illustrating that our pipeline handles the extreme articulations and unusual viewpoints typical of athletic motion.



Figure 8: **SMART Poses Generated.** Examples of images generated using poses from the SMART dataset.

Table 8: **SMART Comparison.** Per-action PCK (↑) against Diffusion-HPC across six athletic actions. "Extra Cond." indicates whether the method requires additional conditioning beyond pose: Diffusion-HPC uses real SMART images as image-to-image starting points, while PoseDreamer generates from scratch conditioned only on pose.

Method	Extra Cond.	Diving	Pole Vault	High Jump	Uneven Bars	Balance Beam	Vault
Diffusion-HPC	✓	79.2	**77.7**	78.1	44.1	85.1	**66.9**
PoseDreamer	×	**81.7**	73.7	**81.7**	**45.1**	**85.4**	65.1

A.3 Spatial Control Ablations

To train a reliable control mechanism for dataset sample generation, we experimented with different representations of the control image. Our initial attempt involved generating image–label pairs by running a pretrained DensePose [15] model on a subset of LAION [51] images. However, these predictions were often inaccurate or incomplete, or had low-quality boundaries, resulting in poor pixel alignment for our control mechanism.

We next considered rendering DensePoseCOCO [15] and AGORA [45] ground truth annotations to ensure high-quality image pairs. Initially, we rendered SMPL-X parameters using the colormap from Continuous Surface Embeddings [42]. This colormap, however, maps each vertex to a value between 0 and 1, which provides insufficient visual variation across poses, particularly for distinguishing head orientations or body configurations. Moreover, this colormap is only defined for SMPL parameters, whereas our goal is to model the full body pose using the SMPL-X model [46]. We were able to obtain a stable and reliable control using the PNCC colormap [71], and mapping each X,Y,Z normalized coordinate to R,G,B values. Following the ablations from the main paper, we compare different representation choices for the conditioning image and report their impact on downstream model performance in Table 9. Using high-quality image pairs and a more detailed colormap significantly improves downstream model performance.

Table 9: **Control Representations.** We compare the effect of different representation choices for the control model training. The comparison is based on downstream model performance.

Control 2D Representation	UBody PVE ↓	Egobody PVE ↓	EHF PVE ↓
PoseDreamer (PNCC)	**104.43**	**128.02**	**95.08**
CSE [42] colormap	129.83	130.66	130.71
Densepose predictions	119.95	153.10	151.40

A.4 Filtering Statistics

To ensure the quality of the generated dataset, we introduce a filtering pipeline. We provide additional statistics on the scores of the generated images. Figure 10 shows the histogram of the OKS score distribution for pose predictions on 60,000 generated images, along with examples of low- and high-scoring images. In addition, we present the histogram of the distribution of mean absolute errors (MAE) for roll, pitch, and yaw in the 3D head pose parameter predictions. The thresholds used for the final dataset selection are 0.8 for the OKS score and 25° for the head pose error.

Pose distribution and pass rates. To verify that filtering does not systematically discard valid-but-hard poses, we compare pre- and post-filter pose distributions in Figure 9 (t-SNE over SMPL-X body-pose parameters). Filtering preserves the overall pose space, confirmed by the low JS divergence between pre- and post-filter distributions (0.031 for LAION+AMASS, 0.042 for MOYO), while BEDLAM and MOYO occupy subregions of our broader distribution. Pass rates differ by source; for instance, the OKS filter pass rate is 77.7% for LAION+AMASS and 60.1% for MOYO, consistent with MOYO's out-of-domain yoga poses (neither the control model nor DPO was trained on yoga).

Threshold sensitivity. Sweeping OKS (0.7–0.9) and head-pose MAE (15°–25°) thresholds over 400K generated images changes downstream performance by at most ∼3%. Best values depend on training setup and evaluation set; we use 0.8 (OKS) and 25° (MAE) as a balanced overall recipe.

Distribution balancing. We balance gender and ethnicity when sampling identity descriptors during generation. Other distributions (e.g., lighting or activity mix) are task-dependent and can be adjusted through prompting without retraining the pipeline.

A.5 Image Quality

To quantitatively evaluate the synthetic-to-real domain gap, we compute the quality and coverage of samples produced by a generative model following [30], and compare them to the faces subset of the LAION dataset [51]. The results are shown in Figure 11. We observe that the generated images closely match the real distribution, whereas rendered datasets achieve a much lower F-score. Furthermore, UMAP [39] and t-SNE [59] projections of DINO-v3 [53] and CLIP [47] image features show that PoseDreamer samples occupy the same region of the real data manifold. In comparison, BEDLAM [5], AGORA [45], and SynBody [67]

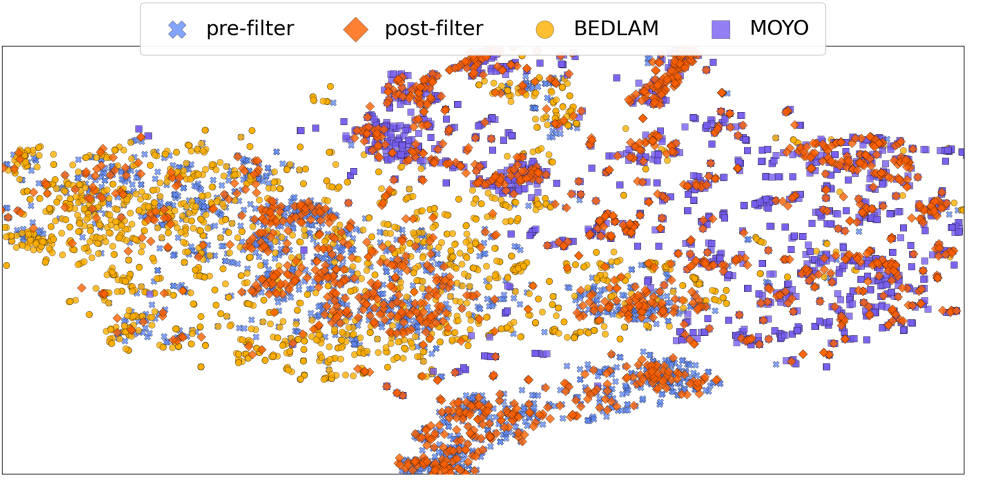


Figure 9: **Pose distribution.** Filtering preserves the overall pose space; BEDLAM and MOYO occupy subregions of PoseDreamer's distribution.

form isolated clusters. Table 10 shows the corresponding Inception Score and FID metrics: PoseDreamer substantially outperforms rendered synthetic datasets on both.

Table 10: **Image Quality:** Diffusion-generated data achieves superior realism than rendering-based synthetic datasets.

Dataset	Inception Score ↑	FID ↓
PoseDreamer	**9.78**	**1.72**
BEDLAM [6]	4.35	4.55
SynBody [57]	5.39	5.19
AGORA [45]	5.55	8.14

A.6 2D Reprojection

Our synthetic dataset's precise ground truth annotations enable the incorporation of additional explicit supervision signals within the SMPL-X topology. We explore how rich synthetic data can improve mesh recovery through enhanced architectural supervision. We introduce a minor modification to the SMPLer-X architecture by adding intermediate visual marker predictions. While SMPLer-X originally predicts 3D joint locations as input to pose parameter heads, we extend this concept by incorporating visual markers as a sparse set of 64 fixed vertices on the mesh surface, following VirtualMarker [36]. Unlike joints, which primarily capture pose information, visual markers are more sensitive to shape variations, providing valuable signals for predicting shape parameters.

This approach provides shape-sensitive supervision while maintaining efficiency through sparse representation. Our modified architecture predicts both joints and visual markers simultaneously, feeding marker predictions to the shape parameter head as additional input. We apply supervision losses to markers in both 3D and 2D spaces, directly on predicted markers and after reconstructing them from predicted avatars, analogous to joint supervi-

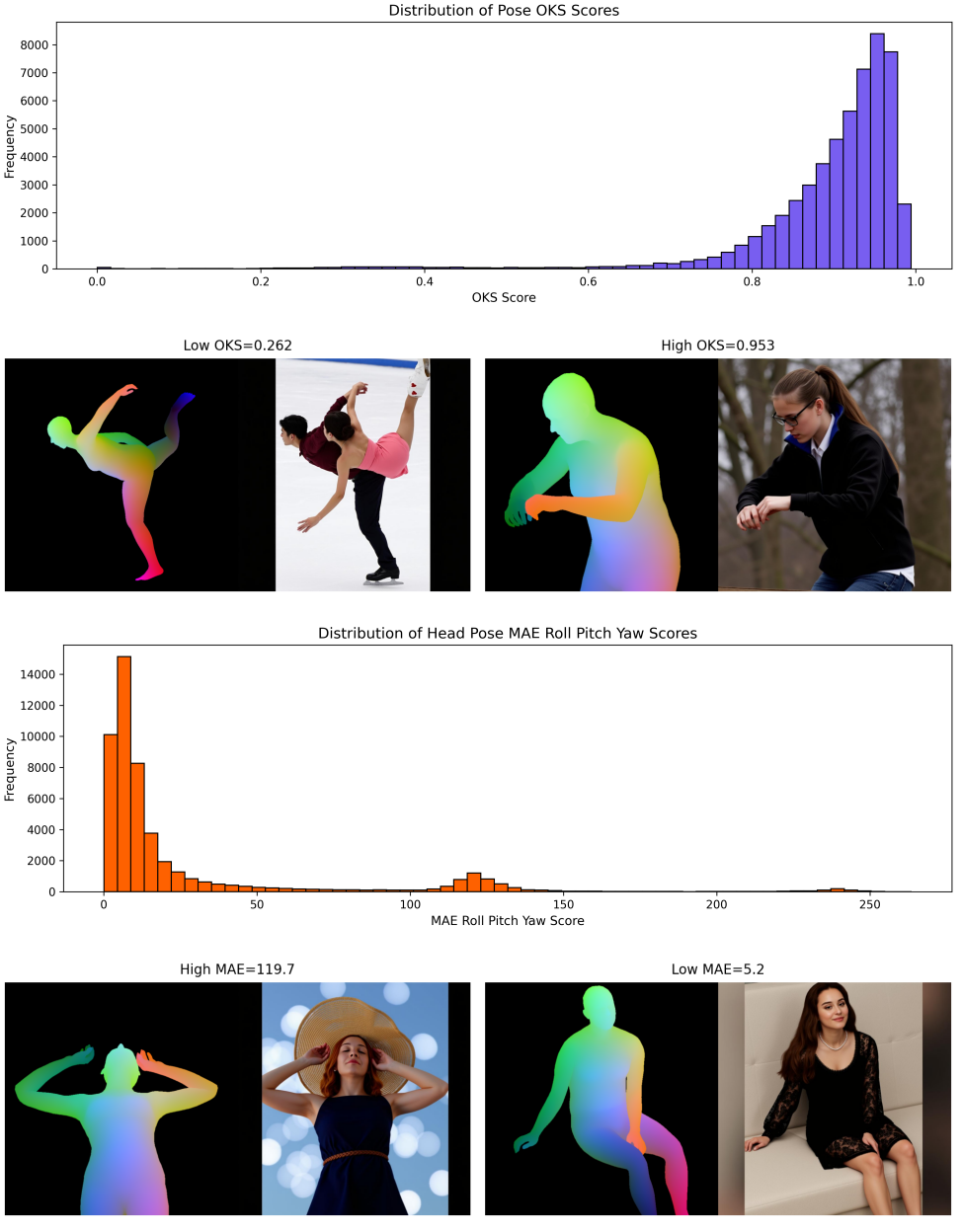


Figure 10: **Filtering statistics.** Distribution of the pose OKS errors, and head pose MAE for roll pitch yaw, together with examples of low- and high-scoring images.

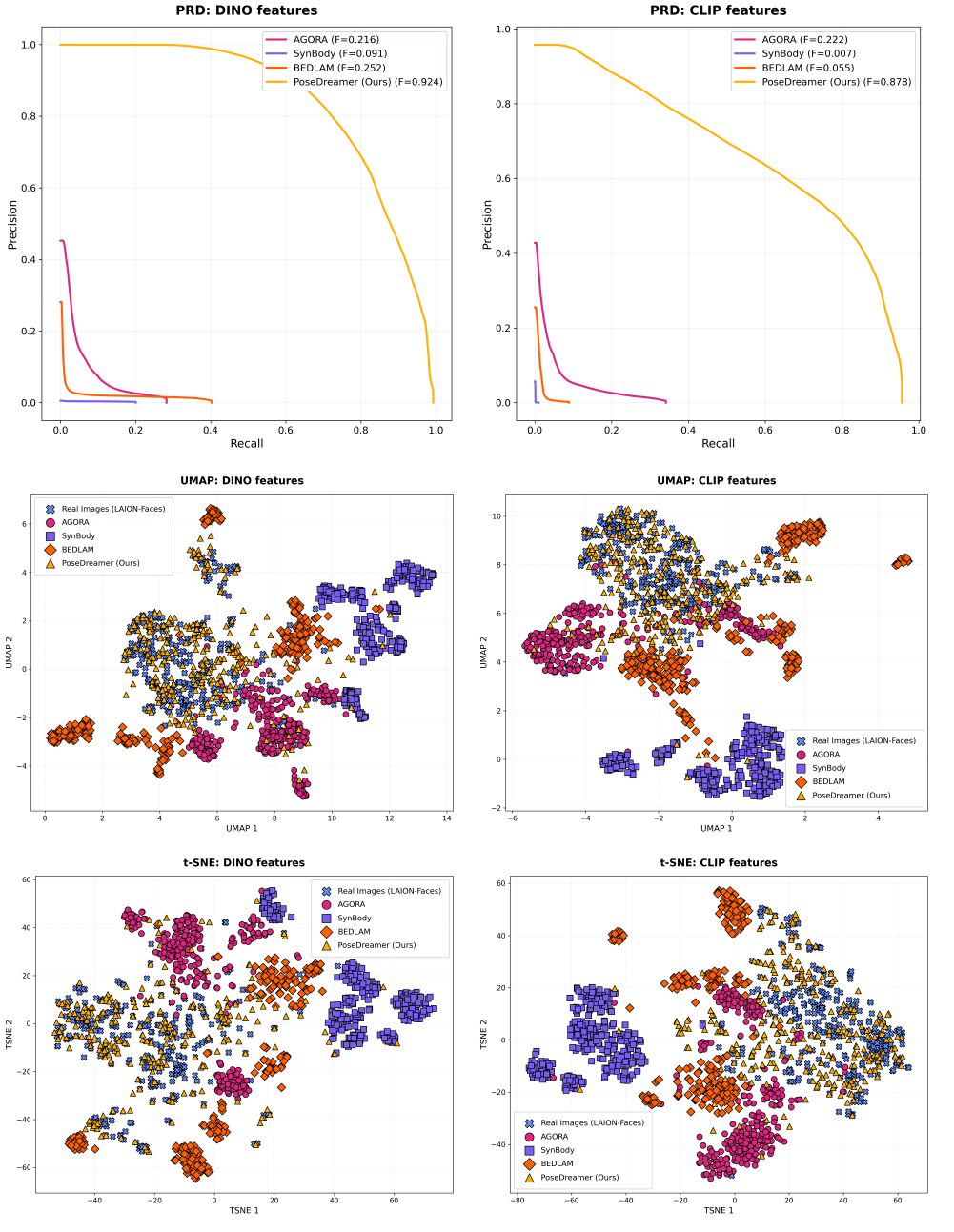


Figure 11: **Image Quality Analysis.** Results for quality and coverage of the samples following [30]. UMAP [39] and TSNE [59] projections of DINO [53] and CLIP [47] image features for PoseDreamer compared to real images from the LAION faces dataset and other synthetic datasets. Samples from our generated dataset more closely follow the distribution of the real dataset.

sion in the original architecture. Table 11 demonstrates the effectiveness of this enhanced supervision. Our model with visual marker supervision achieves 2D reprojection errors approaching SMPLer-X trained on 32 combined real and synthetic datasets, demonstrating that high-quality synthetic supervision can achieve competitive performance with significantly fewer data sources.

Table 11: **2D Reprojection Error Analysis.** PoseDreamer provides rich supervision via visual markers, achieving 2D reprojection accuracy approaching that of models trained on 32 datasets while using only our single synthetic source.

Training Data	EHF 2D Reproj ↓	EgoBody 2D Reproj ↓	UBody 2D Reproj ↓
32 Datasets	**0.013**	0.026	**0.024**
PoseDreamer	0.017	0.027	0.030
PoseDreamer + Markers	0.015	**0.024**	0.025



Figure 12: **PoseDreamer Failure cases.** There are some failure cases for our method. These include unrealistic body shapes for complex input poses or input descriptions that involve complex interactions with other people. Additionally, it can happen that the model fails to understand the 3D pose perspective of the input, or that it places the person in an unrealistic scenario (for example, floating with no support).

B Failure Cases

While our method is effective, it has limitations: the generated images are not always fully photorealistic or perfectly aligned with the control input. Most mistakes are removed during filtering, but minor ones occasionally slip through. Manual inspection of 100 random samples found 3 hand-detail artifacts (3%) and no face artifacts. These match known FLUX limitations rather than DPO-induced failures: DPO's OKS-based reward has blind spots, as anatomical errors that preserve keypoint positions do not affect OKS and pass through. Region-specific hand and face losses are therefore a natural future direction.

The failures that do survive fall into three groups. Complex body poses, such as yoga positions or intricate multi-person interactions described in the text, can produce deformed

body parts or additional limbs. Physically implausible scenarios arise, such as people appearing to float. Finally, the model sometimes struggles with the 3D-to-2D mapping of limb joints, bending legs in the generated image that are straight in the control input as it incorrectly adjusts them for image perspective. Examples are shown in Figure 12.



Figure 13: **Effectiveness of Realism Adapter.** We show comparisons of the same image generated without (left) and with the realism LoRA adapter. Images generated without it tend to be more oversaturated, giving an "AI look".

C Mitigating Distribution Gap in Diffusion-Generated Data

Diffusion models trained on large-scale web datasets like LAION often exhibit aesthetic biases, resulting in oversaturated, heavily filtered imagery with an artificial "AI look." This happens because these models learn to copy the visual styles common in retouched internet images, which are quite different from natural image distributions, leading the generated synthetic data to look unrealistic.

To address this distribution gap, we use a specialized LoRA adapter [18] in our pipeline, trained on real-world images to reduce aesthetic artifacts from foundation models. This adapter learns to transform images from the biased aesthetic style of web-trained diffusion

models toward more natural image appearance. The modular design of EasyControl [70] allows to combine the aesthetic/style adapter with our spatial control LoRA at inference time, producing images that maintain both precise spatial consistency with control inputs and improved realistic appearance.

Figure 13 demonstrates the effectiveness of this approach, showing substantial reduction in oversaturation and artificial enhancement artifacts. This aesthetic correction represents a crucial step in bridging the domain gap between synthetic training data and real-world deployment scenarios.

D Control Model Alignment: Implementation Details

For DPO training, we generate 40,000 preference pairs by creating 4 image variants for each caption and 3D mesh combination, then ranking them based on OKS scores to establish preferred and less preferred samples. We use a lower β value of 250 compared to the original Flow-DPO [34] value of 500, as we find that control models require less regularization due to their more constrained generation space. Additionally, traditional ControlNet approaches may face challenges with RL-based alignment, as both the control mechanism and alignment procedures operate on diffusion model activations, potentially leading to optimization conflicts that can compromise training stability. In contrast, EasyControl [70] modular adapter architecture we use provides cleaner separation between control and alignment components, facilitating more stable joint optimization.

E Additional Visualizations

We provide some additional examples of dataset samples in Figure 14.

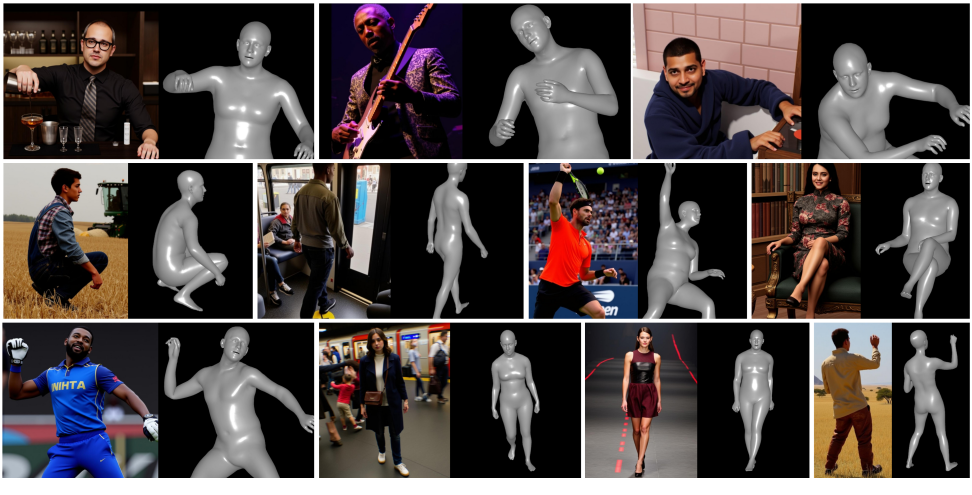


Figure 14: **PoseDreamer Dataset Samples.** Examples from PoseDreamer dataset showing diverse poses, scenes, and high-quality image generation with precise 3D mesh correspondence.