Avatar4D: Synthesizing Domain-Specific 4D Humans for Real-World Pose Estimation

Anonymous CVPR submission

Paper ID *****

Abstract

We present `Avatar4D`*, a real-world transferable pipeline for generating customizable synthetic human motion datasets tailored to domain-specific applications. Unlike prior works, which focus on general, everyday motions and offer limited flexibility, our approach provides fine-grained control over body pose, appearance, camera viewpoint, and environmental context, without requiring any manual annotations. To validate the impact of* `Avatar4D`*, we focus on sports, where domain-specific human actions and movement patterns pose unique challenges for motion understanding. In this setting, we introduce* `Syn2Sport`*, a large-scale synthetic dataset spanning sports, including baseball and ice hockey.* `Syn2Sport` *features high-fidelity 4D (3D geometry over time) human motion sequences with varying player appearances rendered in diverse environments. We benchmark several state-of-the-art pose estimation models on* `Syn2Sport` *and demonstrate their effectiveness for supervised learning, zero-shot transfer to real-world data, and generalization across sports. Furthermore, we evaluate how closely the generated synthetic data aligns with real-world datasets in feature space. Our results highlight the potential of such systems to generate scalable, controllable, and transferable human datasets for diverse domain-specific tasks without relying on domain-specific real data.*

1. Introduction

Understanding human motion is central to a wide range of vision tasks, from activity recognition and biomechanics to sports analytics [7, 11] and interaction modeling [15, 36]. General-purpose human pose datasets [1, 19, 26, 43] provide broad coverage of everyday activities, but they lack domain-specific motions and rarely include challenging poses or sport-specific dynamics. Existing sports-focused datasets [2, 6, 11, 21, 39] offer more targeted coverage but remain limited to popular sports such as soccer, leaving many activities and their unique motions underrepresented, such as skating in ice hockey or pitching in baseball. This gap motivates the need for domain-specific large-scale datasets that reflect their diversity and complexity.

Synthetic data offers a promising alternative, providing scalable access to diverse and precisely labeled samples. While recent synthetic datasets [5, 9, 30, 32, 42] have enabled progress in modeling general everyday human activities, they do not capture domain-specific challenges such as high acceleration, unusual joint configuration, or unusual temporal dynamics. In addition, they lack fine-grained control over body posture, camera viewing perspective, appearance variability (such as clothing, body shape, and textures), lighting, and background context, which we refer to as the *configurable simulation parameters*.

This raises a key challenge: *How can we generate synthetic human datasets that are configurable to domain requirements and effective enough to substitute for real-world data in those domains?*

Motivated by this question, we introduce `Avatar4D`, a generation pipeline for synthetic human motion datasets tailored to specific domains. The key challenge is to capture diverse, unbiased distributions that still transfer well to real data. As illustrated in Figure 1, our pipeline operates in three stages: (1) we build a dictionary of domain-relevant inputs, including source person images, motion sequences from internet videos, and background scenes; (2) each person image is lifted into a canonical 3D asset, which is then animated by deforming it according to the chosen motion sequence; and (3) the animated 3D human is rendered against the corresponding background to form a realistic frame. The final output is a large-scale collection of high-fidelity 4D human sequences (spanning diverse identities, motions, and environments) that can be *directly leveraged for training and evaluation in real-world conditions.* For example, `Avatar4D` can generate basketball players dribbling on a court, construction workers moving at a site, or pedestrians navigating streets, enabling domain-specific benchmarks for human-centric tasks such as pose estimation, 3D human modeling, and motion segmentation.

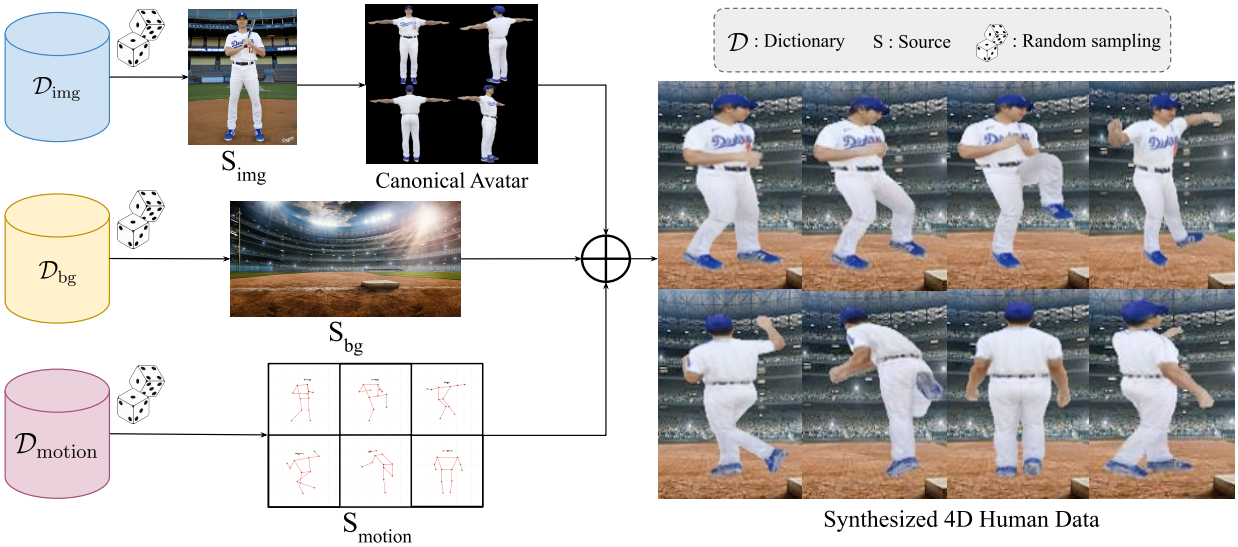

Figure 1. **Overview of Avatar4D architecture.** We sample a random source image (S_{img}), source background (S_{bg}) and motion sequence (S_{motion}) from three dictionaries: $\mathcal{D}_{img}$ for domain-relevant person images, $\mathcal{D}_{bg}$ for scene/background images, and $\mathcal{D}_{motion}$ for motion sequences. These inputs are then combined to generate domain-specific synthetic 4D human data.

To demonstrate and validate the capabilities of `Avatar4D` in generating realistic, domain-specific synthetic data, we use it to create `Syn2Sport`, a large-scale synthetic dataset spanning sports domains, including baseball and ice hockey. We specifically focused on these sports because of their distinct motion patterns: fine-grained single-athlete actions in baseball and fast-paced skating action in ice hockey. `Syn2Sport` contains high-fidelity human motion sequences rendered across diverse environments with rich appearance variations, providing a controlled testbed to evaluate the effectiveness of the proposed `Avatar4D` approach. Importantly, to address our central research question, we conducted extensive experiments using the `Syn2Sport` dataset, including zero-shot synthetic-to-real pose transfer, domain generalizability, and feature-space alignment. The results demonstrate that `Avatar4D` enables the generation of synthetic datasets that are both configurable to domain-specific requirements and transferable to real data, thereby reducing reliance on real-world labels. Our contributions can be summarized as follows:

1. We propose `Avatar4D`, a real-world transferable synthetic pipeline for generating domain-specific datasets, with precise control over the configurable simulation parameters.
2. We introduce `Syn2Sport`, a large-scale synthetic dataset featuring high-fidelity 4D human motion across sports domains (baseball and icehockey).
3. We conduct extensive experiments on zero-shot domain generalization, synthetic-to-real transfer across different sports, and feature-space alignment.

2. Related Works

2.1. Real Datasets

Real-world datasets have been foundational in advancing Human Pose Estimation (HPE). HPE aims to extract the 2D human pose of a person given a single image. Some commonly benchmarked datasets include COCO [26] and MPII [1], which provide large-scale annotated images of everyday activities, supporting robust keypoint detection. There are some other task-specific datasets, such as sports-focused datasets like LSP [21] offer annotated images for athletic movements, while datasets such as PoseTrack [12], OCHuman [48], and CrowdPose [24] capture multi-person interactions with heavy occlusions.

While these datasets have driven significant progress in 2D HPE, they exhibit several limitations: (1) Most datasets focus on everyday activities, offering limited coverage of specialized domains such as sports, dance, or rehabilitation tasks; (2) Rare or complex poses are underrepresented, which restricts model generalization; and (3) Capturing and labeling 2D pose data is time-consuming and are prone to human annotation bias. These limitations with existing real-world datasets make it necessary to develop alternative approaches to collect domain-specific training data without having to manually label it.

2.2. Synthetic Datasets

Synthetic datasets provide a scalable and flexible alternative by leveraging 3D modeling, motion capture, and rendering techniques. Notable examples include SURREAL [42], AGORA [30], BEDLAM [5], and 3DPeople [32], which provide annotated 2D and 3D poses with synthetic images.

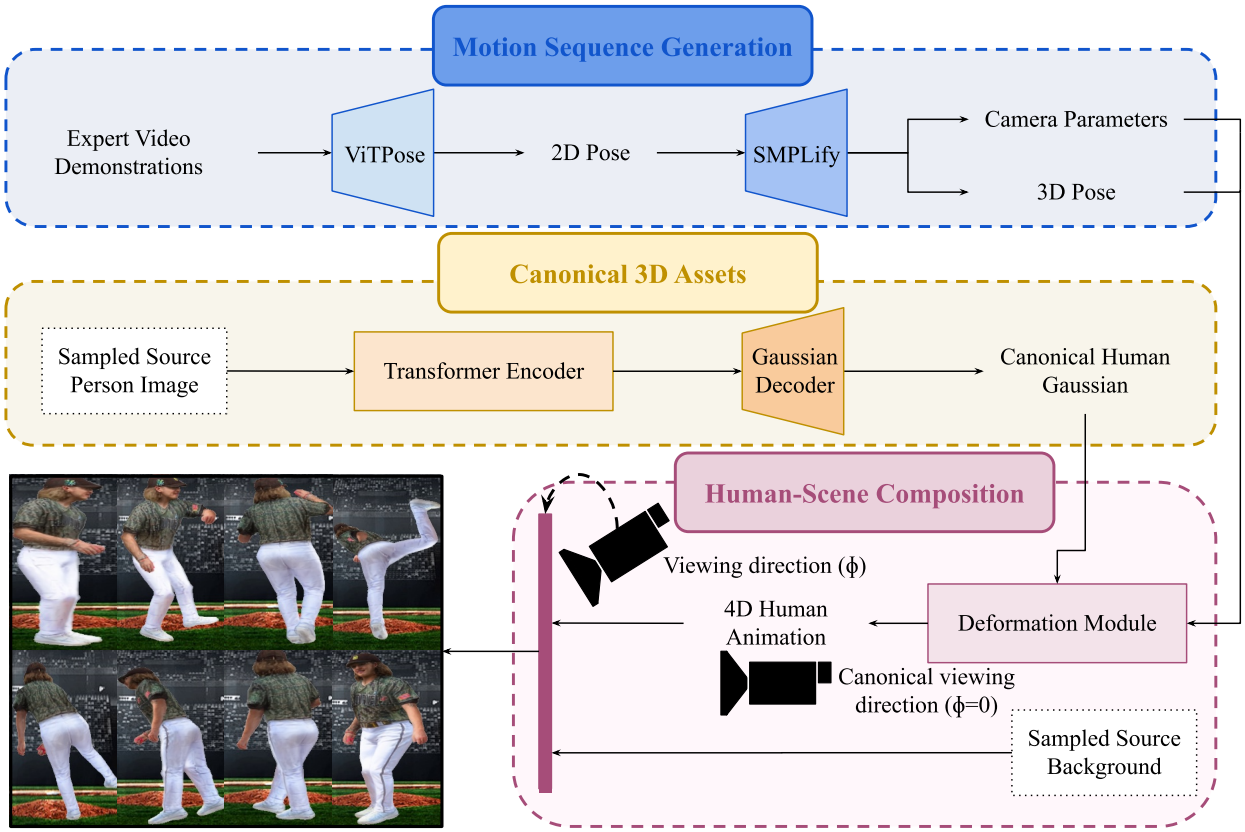

Figure 2. **Architecture of Avatar4D.** The proposed pipeline consists of three key stages. First, a motion sequence is constructed from expert demonstrations collected from online sources, producing 3D poses along with corresponding camera parameters. Next, canonical 3D human assets are generated from sampled source person images. Finally, in the human-scene composition stage, these assets are deformed according to the 3D poses and rendered against sampled source background.

More recently, methods such as Gen4D [8] allow dynamic 4D human synthesis with fine-grained control over body posture, appearance, and background; however, it doesn't transfer well to real-world scenarios.

However, existing synthetic datasets still have important limitations: (1) Most synthetic datasets emphasize everyday actions, lacking domain-specific coverage for tasks like sports analytics, rehabilitation, or industrial applications; (2) Artifacts introduced by rendering pipelines or background generation can reduce realism, impeding direct transfer to real-world models; and (3) Restricted fine-grained control over pose sequences, camera viewpoints, or environmental context simultaneously, which is crucial for generating task-specific training data.

These gaps highlight the need for a domain-adaptive, high-fidelity synthetic pipeline that can generate realistic 4D human motion data for target domains while ensuring transferability to real-world scenarios.

3. Methodology

3.1. Overview

The goal of the proposed `Avatar4D` generation pipeline is to synthesize a large-scale 4D-human sequence with controllable motion, realistic geometry, and photorealistic rendering. To achieve this, we decouple the problem into three stages: motion generation, asset creation, and sequence rendering. So we begin with controllable motion data generation, which is then paired with canonical 3D asset generation from monocular human images. The generated motion is used to deform the 3D asset, and the resulting sequence is rasterized against a static background. Figure 2 illustrates the overall pipeline, while subsequent sections detail each component of the pipeline.

3.2. Motion Sequence Generation

Prior works [5, 42] typically utilize motion sequences derived from MoCap datasets [29, 41], which are commonly used to drive the deformation of 3D human assets. However, these datasets are often limited in diversity and fail to capture the breadth of motion found in in-the-wild, domain-

specific scenarios, such as those encountered in sports or street activities. To address this limitation, we construct a dictionary of motion sequences that better reflects real-world variability. Unlike recent generative methods that synthesize motion from text prompts using motion diffusion models [3, 35, 40], we curate motion data directly from internet videos of target activities, allowing us to ground the motion in real-world dynamics.

Specifically, we extract whole-body 2D keypoints (pseudo) from these videos using a pretrained pose estimator [44]. These pseudo keypoints are then lifted to 3D body parameters by fitting an SMPL-X mesh using SMPLify-X [31]. While the extracted pseudo keypoints may be noisy, they are only used to drive synthetic avatar generation. This ensures that the final generated avatars still provide perfect image–pose pairs under controllable camera viewpoints. The resulting sequences serve as pseudo-ground truth to guide the 3D assets for downstream composition.

3.3. Canonical 3D Assets

3D asset generation has traditionally relied on CGI pipelines and animation experts. This manual process is time-consuming and inherently limits variability, especially in task-specific settings where asset appearance (e.g., clothing, body shape, accessories) needs to adapt to the domain. A promising alternative is the use of 3D Gaussian Splatting (3DGS) [22], which allows for fast and high-fidelity rendering of dynamic or clothed human avatars without explicit mesh reconstruction.

Recent advances in 3DGS for humans have explored asset creation from various input modalities, ranging from multi-view images [17, 23, 33], text-guided templates [14, 27, 28], to single-view images [34, 49]. In our pipeline, we adopt the LHM [34] method, which enables reconstruction of animatable 3D avatars from a single input image. Compared to multi-view reconstruction methods [17, 33], which require cumbersome data collection and calibration, or text-prompt-driven methods [14, 27], which often generate avatars with limited realism and domain fidelity, LHM offers a practical solution for scalable dataset generation. Each avatar is represented as a collection of 3D Gaussians, where each Gaussian is parameterized by a centroid $p \in \mathbb{R}^3$, an anisotropic scale $\sigma \in \mathbb{R}^3$, and a quaternion $r \in \mathbb{R}^4$ that encodes its rotation. Additionally, an opacity value $\rho \in [0, 1]$ controls the transparency of the Gaussian, while a feature vector $f \in \mathbb{R}^C$, encoded using spherical harmonics, captures view-dependent appearances. These parameters collectively form the tuple $\chi = (p, r, f, \rho, \sigma)$, which defines the full set of properties for each point in the 3DGS representation.

3.4. Human-Scene Composition

3.4.1. View Space Rendering

Given the canonical 3DGS asset parameters (χ), i.e., the canonical (undeformed rest-pose) representation of the avatar in its reference space, we first transform the canonical avatar to target view space using Linear Blend Skinning (LBS) based on the motion sequence. Each Gaussian ($\mathbf{x}_i$) is transformed as shown in Equation (1).

$$\mathbf{x}_{i'} = \sum_{j=1}^{J} w_{ij} \mathbf{T}_j \mathbf{x}_i \tag{1}$$

where w_{ij} are the skinning weights and $\mathbf{T}_j \in \mathbb{R}^{4\times4}$ are the per-joint transformation matrices computed via forward kinematics. The deformed Gaussians are then rendered through differentiable splatting under the target camera parameters π_t to produce the RGB image $\hat{I}$ and alpha mask $\hat{M}$. The final composite image $\mathcal{I}$ is obtained by alpha-blending the rasterized foreground with the background image I_{bg}:

$$\mathcal{I} = \hat{I} \odot \hat{\mathbf{M}} + I_{\text{bg}} \odot (1 - \hat{\mathbf{M}})$$

where I_{bg} is the background image, which is alpha-blended with the rasterized foreground.

Keypoints Transformation. The 2D keypoints generated from Section 3.2 must be transformed to align spatially with the rendered view space. This transformation includes three sequential steps: centering concerning the original camera principal point, scaling to match the canonical render resolution r, and repositioning within the final image frame. Formally, given an original keypoint (x, y) and its associated principal point $(c_x^{\text{old}}, c_y^{\text{old}})$, the transformed keypoint coordinates (x', y') in the rendered image space of dimensions $(w_{\text{new}}, h_{\text{new}})$ are computed as:

$$x' = (x - c_x^{\text{old}}) \cdot \frac{r}{w_{\text{new}}} + \frac{w_{\text{new}}}{2} \tag{2}$$

$$y' = (y - c_y^{\text{old}}) \cdot \frac{r}{h_{\text{new}}} + \frac{h_{\text{new}}}{2} \tag{3}$$

4. Dataset

We present `Syn2Sport`, a large-scale synthetic human motion dataset covering sports such as baseball and ice hockey, generated using the `Avatar4D` pipeline. `Syn2Sport` comprises 21,203 clips across all splits, totaling roughly 4.48 million frames and over 2,400 minutes of play time after augmentation. Some sample images from our dataset are illustrated in Figure 3.

The dataset is split into training and validation sets per sport, as summarized in Table 1. Each sample includes 2D

Figure 3. **Sample images from the Syn2Sport dataset.** All samples were generated using the proposed `Avatar4D` pipeline.

pose annotations in COCO-WB [20] format, 3D pose annotations, and full 3D human pose and shape parameters represented in the SMPL-X model. This combination of 2D and 3D annotations enables training and evaluation across a broad range of human pose estimation tasks, from 2D keypoint detection to detailed 3D human mesh recovery.

Augmentations. To enhance viewpoint diversity in `Syn2Sport`, we implement stochastic camera sampling for each motion sequence. Specifically, the camera's extrinsic parameters (3D position relative to the subject and its orientation), defined by elevation and azimuth angles, are randomly sampled per sequence. This produces varying perspectives capturing the human from different angles.

5. Experiments

5.1. Evaluation

Datasets. We evaluate the performance of the models trained with the data generated from `Syn2Sport` with two real-world datasets. MLBPitchDB [6], which is a dataset focusing on baseball players containing high motion blur and self-occlusion. For ice hockey, we use a large-scale dataset [2] which is annotated from NHL broadcast videos. Furthermore, for zero-shot experimentation across other sports domains, we use two real-world datasets: a soccer dataset [45] and a lacrosse dataset [2]. To provide a baseline comparison against models trained on large-scale real-world data, we also use COCO-WholeBody (COCO-WB) [20], an extension of COCO [26] with 133 keypoints covering the face, body, hands, and feet. For our experiments, we specifically use the first 23 keypoints corresponding to the body and feet. Additionally, we compare against SportPAL [8], a synthetic sports benchmark that most closely aligns with our setting by reconstructing controllable human motions within contextual backgrounds.

Table 1. **Syn2Sport Data Split.** Data statistics across the baseball and ice hockey sports.

Sport	Split	#Subjects	#Clips	#Frames	Play Time (mins)
Baseball	Train	15	1,017	585,000	319
	Valid	5	254	146,413	87
Ice hockey	Train	15	15,946	3,002,000	1,608
	Valid	5	3,986	750,569	477
Total	-	**40**	**21,203**	**4,483,982**	**2,491**

Table 2. **Benchmarking Syn2Sport on 2D pose estimation models.** The best result is highlighted in **bold** format.

Method	Baseball			Icehockey		
	AP^5↑	AP^{10}↑	AP^{15}↑	AP^5↑	AP^{10}↑	AP^{15}↑
DETR [10]	74.5	78.1	82.2	31.7	41.4	48.6
HRNet [38]	83.6	83.6	96.0	43.0	64.4	74.1
UDP [18]	85.2	88.5	92.6	38.9	61.9	75.3
ViTPose [44]	83.6	95.7	97.9	**55.3**	**82.4**	**91.0**
TokenPose [13]	**89.1**	**97.5**	**98.7**	54.2	81.1	90.4

Metrics. To evaluate pose estimation performance, we report the Average Precision (AP) at multiple thresholds. AP measures the percentage of predicted keypoints falling within a specified distance (in pixels) from the ground truth, normalized by the subject's bounding box size. For most experiments, we use thresholds of 5, 10, and 15 pixels. For the more challenging zero-shot sports transfer setting, we adopt slightly relaxed thresholds of 15, 20, and 25 pixels to account for the increased domain gap.

To measure the similarity between the real and synthetic data distributions, we computed the Fréchet Inception Distance (FID) [16] and Kernel Inception Distance (KID) [4] between the deep feature embeddings of the synthetic and real data. FID measures the distance between the Gaussian approximations of the feature distributions, while KID computes an unbiased squared Maximum Mean Discrepancy (MMD) between the distributions.

5.2. Main Experiments

Benchmarking on Syn2Sport Dataset. Since `Syn2Sport` provides large-scale synthetic sports sequences with ground truth 2D pose annotations, we use it as a benchmark to assess the quality of synthetic data for HPE. Specifically, Table 2 evaluates state-of-the-art 2D HPE models on the baseball and icehockey subsets using keypoint accuracy (AP), thereby quantifying how well these models perform when applied to these domains.

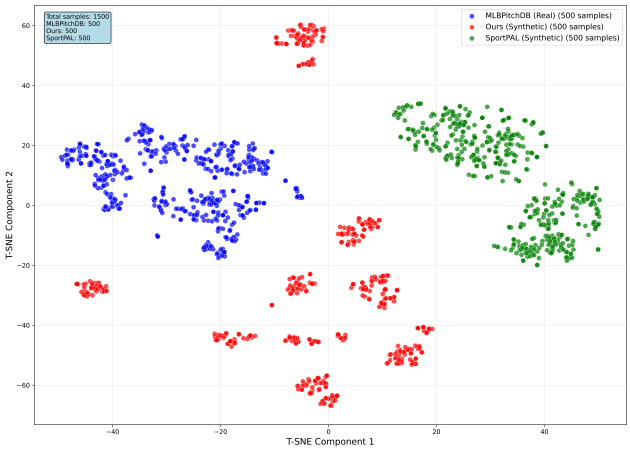

Figure 4. **t-SNE visualization of feature embeddings.** Real samples (red) and synthetic samples (blue) show substantial overlap, indicating strong similarity between the datasets.

(a) ControlNet (b) IC-Light (c) Static

Figure 5. **Background generation using different methods.** Samples generated using ICLight, ControlNet, and static backgrounds, illustrating their impact on the realism of the synthetic human data.

Zero-Shot Synthetic-to-Real Transfer. The ultimate goal of the proposed method and its resulting dataset is to generate synthetic data that closely represents the real domain. To evaluate this, we analyze how models trained on the synthetic data transfer to real-world datasets. We consider three training configurations: training on the COCO-WB dataset alone, training on the synthetic dataset alone, and training on both datasets combined. Performance is measured on two real-world sports datasets: baseball [6] and ice hockey [2], with all models trained using TokenPose [25].

The results, as shown in Table 3 demonstrate that incorporating synthetic data significantly improves performance. Specifically, adding synthetic data yields 55.96% and 114.5% improvements in AP^5 on the real-world baseball and ice hockey datasets, respectively, compared to training on COCO-WB alone. While COCO-WB contains some real baseball examples and slightly outperforms the synthetic-only model on MLBPitchDB, combining COCO-WB with synthetic data produces the best results. In contrast, COCO-WB alone fails to generalize to ice hockey due to the absence of skating-related samples, where the synthetic-only model achieves substantial improvements. These results demonstrate that *the synthetic data generated from* `Avatar4D` *can effectively substitute for real-world, domain-specific training samples*, enabling models to be pretrained on general-purpose datasets and then finetuned solely on our synthetic samples for robust results.

Figure 6 shows examples of pose transfer from our synthetic dataset to real-world images. Models trained on `Syn2Sport` accurately predict keypoints on real-world baseball and ice hockey datasets, demonstrating improved alignment compared to training solely on COCO-WB.

Evaluation of Synthetic Data Fidelity. To quantitatively and qualitatively assess how closely our synthetic dataset resembles the real data, we conducted a feature-space similarity analysis. First, we extract high-level image features [37] from both the synthetic and real datasets to capture semantic and perceptual image information. Results as shown in Table 4 demonstrate that our synthetic dataset improves FID by 34% and KID by 54%, demonstrating better feature-space alignment compared to SportPAL [8].

To complement the quantitative metric, we performed t-SNE dimensionality reduction on the same feature embeddings, projecting them into a 2D space as shown in Figure 4. t-SNE was computed on feature embeddings with perplexity of 30 and 1000 iterations. In the resulting visualization, real and synthetic samples were collected from MLBPitchDB, SportPAL, and `Syn2Sport`, confirming that our synthetic samples were positioned closer to real data than the data generated from SportPAL, indicating improved semantic and perceptual similarity.

5.3. Ablation Study

Impact of Background on Synthetic-to-Real Transfer. We evaluate the impact of different background generation methods on synthetic-to-real transfer for pose estimation, comparing static image backgrounds, IC-Light [47], and ControlNet [46]. Figure 5 shows examples of backgrounds generated using each approach. Generative methods often alter the human asset itself: IC-Light produces an animation-like, artificial appearance, while ControlNet occasionally introduces spurious artifacts, such as an extra hand (highlighted by the red arrow), which degrade model performance. In contrast, static backgrounds preserve the fidelity of the human model, achieving superior performance as confirmed by the quantitative analysis in Table 5.

Cross-Sport Zero-Shot Generalization. To evaluate domain generalization, we train pose estimation models on synthetic data from a single sport (e.g., baseball or ice

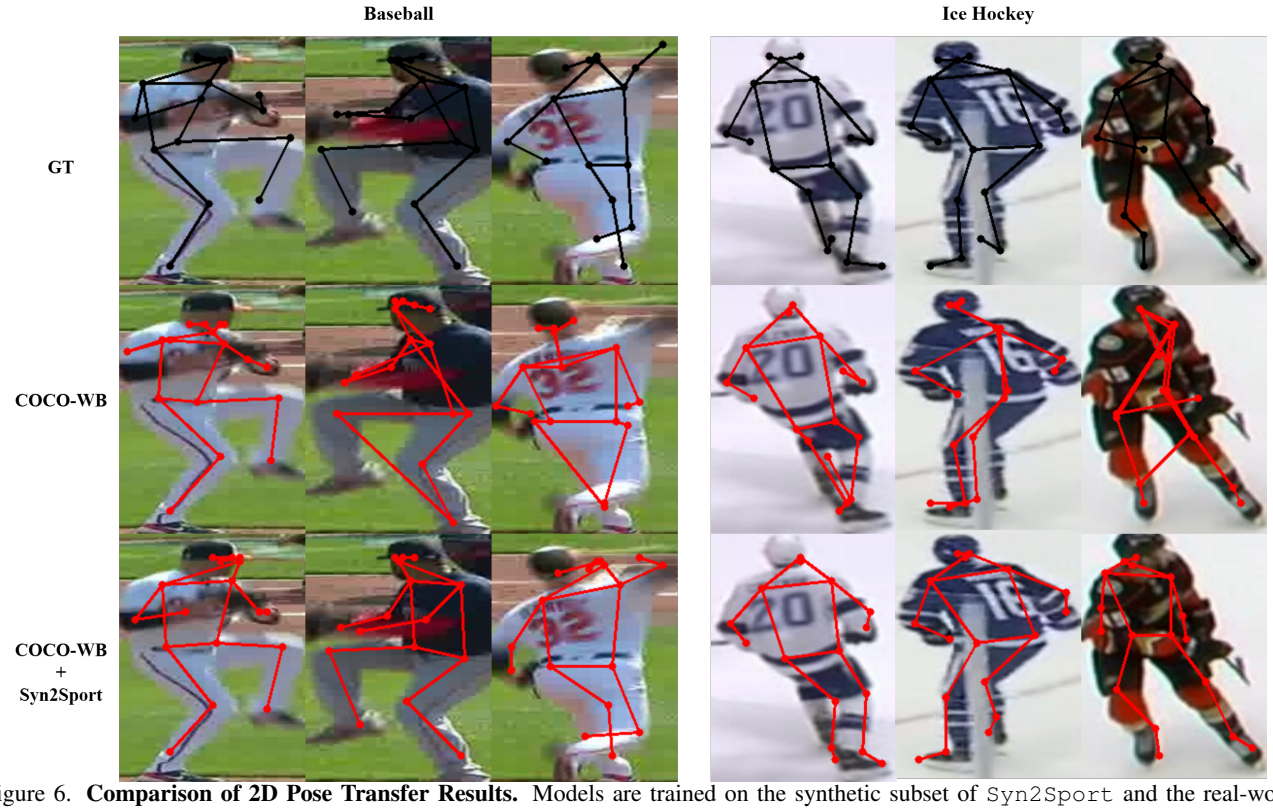

Figure 6. **Comparison of 2D Pose Transfer Results.** Models are trained on the synthetic subset of `Syn2Sport` and the real-world COCO-WB dataset, and evaluated on real-world baseball [6] and ice hockey [2] datasets. COCO-WB+`Syn2Sport` denotes pretraining on COCO-WB followed by finetuning on our synthetic subset, and GT denotes the ground truth from the respective dataset.

Table 3. **2D Pose Transfer Performance** on real-world baseball [6] and ice hockey [2] datasets.

Training Data	Baseball			Ice hockey		
	$AP^{5}\uparrow$	$AP^{10}\uparrow$	$AP^{15}\uparrow$	$AP^{5}\uparrow$	$AP^{10}\uparrow$	$AP^{15}\uparrow$
COCO-WB	29.3	56.0	71.3	13.1	39.5	58.3
`Syn2Sport`	23.0	41.3	55.0	19.5	44.9	62.3
COCO-WB+`Syn2Sport`	**45.7**	**77.6**	**90.7**	**28.1**	**61.2**	**82.0**

Table 4. **FID comparison between synthetic datasets and real data.** Lower values indicate closer alignment with the real data.

Dataset	FID ↓	KID ↓
Real Data (reference)	0	0
SportPAL	21.6	0.09
`Syn2Sport` (Ours)	**14.3**	**0.04**

Table 5. **Impact of Background on 2D Pose Transfer** on the real-world baseball [6] and ice hockey [2] datasets.

Background	Baseball			Ice hockey		
	$AP^{5}\uparrow$	$AP^{10}\uparrow$	$AP^{15}\uparrow$	$AP^{5}\uparrow$	$AP^{10}\uparrow$	$AP^{15}\uparrow$
ControlNet [46]	38.1	67.7	84.0	18.4	43.4	71.9
IC-Light [47]	3.4	11.3	20.2	1.3	6.7	15.4
Static	**45.7**	**77.6**	**90.7**	**21.7**	**52.2**	**74.8**

Table 6. **Zero-Shot Generalization Across Sports Domains** with data trained on `Syn2Sport` dataset.

Train	Test (Real)	$AP^{15}\uparrow$	$AP^{20}\uparrow$	$AP^{25}\uparrow$
COCO-WB (Real)	Lacrosse [2]	64.7	72.2	81.6
Ice hockey (Synthetic)	Lacrosse [2]	66.5	76.8	83.3
COCO-WB (Real)	Soccer [45]	71.3	78.2	83.6
Baseball (Synthetic)	Soccer [45]	67.1	76.1	82.4

hockey) and test them zero-shot on real-world datasets from different but related sports (e.g., soccer and lacrosse, respectively). While some motion patterns or equipment may be shared, such as stick handling in ice hockey and lacrosse, these sports differ in visual context, movement dynamics, and scene composition. This setup allows us to assess whether models trained on synthetic data from one domain can transfer effectively to related domains, highlighting the robustness and versatility of our synthetic dataset.

Table 6 demonstrates strong zero-shot performance, highlighting the model's ability to generalize across both visual and kinematic variations using only synthetic supervision. Notably, the TokenPose model trained on our dataset outperforms on the real-world lacrosse test set and shows comparable performance on the soccer dataset relative to the model trained on the manually annotated COCO-WB dataset. These results underscore the advantage of our syn-

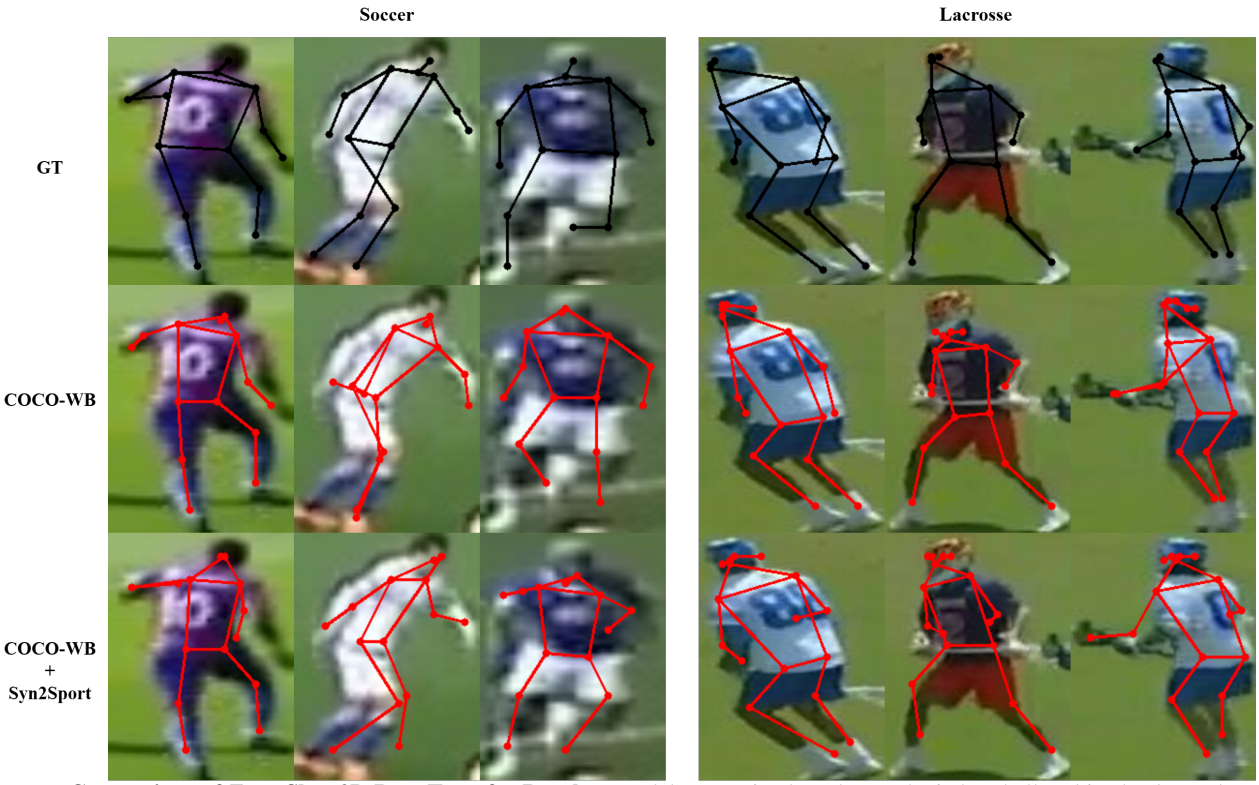

Figure 7. **Comparison of Zero-Shot 2D Pose Transfer Results.** Models are trained on the synthetic baseball and ice hockey subsets of `Syn2Sport` and evaluated on the real-world soccer [45] and lacrosse [2] datasets, respectively.

thetic pipeline, which achieves competitive or superior performance without relying on any manual annotations.

6. Discussion

Our findings provide several practical insights for designing synthetic human motion datasets. First, static backgrounds outperform context-aware generative methods for human-centric motion tasks, as generative backgrounds can introduce an artificial look in the human model that reduces transferability to real data. Second, synthetic data substantially improves real-world performance when combined with models pretrained on COCO, achieving over 50% AP gains on baseball and ice hockey benchmarks. Finally, zero-shot cross-domain generalization is possible, with models trained on one sport transferring to related sports by leveraging shared motion and appearance patterns. Overall, these results highlight that prioritizing human asset fidelity and leveraging synthetic augmentation enables effective, transferable, and task-flexible datasets.

Importantly, the proposed pipeline establishes a domain-adaptive workflow for pose estimation: given a new target domain, a model can be initialized with weights from a pretrained general-purpose dataset (e.g., COCO) and subsequently *fine-tuned exclusively on synthetic data generated from our pipeline*. This eliminates the need for domain-specific real-world annotations while retaining high performance, enabling scalable, task-specific, and transferable human motion modeling.

7. Conclusion

We present `Avatar4D`, a real-world transferable pipeline for generating customizable synthetic human motion datasets for domain-specific applications. Our approach provides fine-grained control over pose, appearance, camera, and environment, enabling the synthesis of high-fidelity 4D human data tailored to target tasks. Using this system, we release `Syn2Sport`, a large-scale synthetic dataset spanning multiple sports, and demonstrate that models trained on it achieve strong supervised performance, zero-shot real-world transfer, and cross-sport generalization. Feature-space analysis confirms that our synthetic data closely aligns with real-world distributions. These results highlight the potential of controllable synthetic data pipelines to reduce reliance on real annotations while enabling scalable and transferable human motion modeling.

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 Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1, 2

[2] Bavesh Balaji, Jerrin Bright, Yuhao Chen, Sirisha Rambhatla, John Zelek, and David Clausi. Seeing beyond the crop: Using language priors for out-of-bounding box keypoint prediction. *Advances in Neural Information Processing Systems*, 37:102897–102918, 2024. 1, 5, 6, 7, 8

[3] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–469, 2024. 4

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5

[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. 1, 2, 3

[6] Jerrin Bright, Yuhao Chen, and John Zelek. Mitigating motion blur for robust 3d baseball player pose modeling for pitch analysis, 2023. 1, 5, 6, 7

[7] Jerrin Bright, Bavesh Balaji, Yuhao Chen, David A Clausi, and John S Zelek. Pitchernet: Powering the moneyball evolution in baseball video analytics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3420–3429, 2024. 1

[8] Jerrin Bright, Zhibo Wang, Yuhao Chen, Sirisha Rambhatla, John Zelek, and David Clausi. Gen4d: Synthesizing humans and scenes in the wild. *arXiv preprint arXiv:2506.05397*, 2025. 3, 5, 6

[9] 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*, 2024. 1

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 5

[11] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3502, 2022. 1

[12] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20963–20972, 2022. 2

[13] 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. 5

[14] Jia Gong, Shenyu Ji, Lin Geng Foo, Kang Chen, Hossein Rahmani, and Jun Liu. Laga: Layered 3d avatar generation and customization via gaussian splatting. *arXiv preprint arXiv:2405.12663*, 2024. 4

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, 2019. 1

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[17] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. 2023. 4

[18] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5700–5709, 2020. 5

[19] 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. 1

[20] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5

[21] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, page 5. Aberystwyth, UK, 2010. 1, 2

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 4

[23] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models, 2023. 4

[24] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 2

[25] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 11313–11322, 2021. 6

[26] 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. 1, 2, 5

[27] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splat-

ting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6646–6657, 2024. 4

[28] Yufei Liu, Junshu Tang, Chu Zheng, Shijie Zhang, Jinkun Hao, Junwei Zhu, and Dongjin Huang. Clothedreamer: Text-guided garment generation with 3d gaussians. *arXiv preprint arXiv:2406.16815*, 2024. 4

[29] 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. 3

[30] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 1, 2

[31] 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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[32] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2242–2251, 2019. 1, 2

[33] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 4

[34] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 4

[35] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 4

[36] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17038–17047, 2023. 1

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 5

[39] Ryota Tanaka, Tomohiro Suzuki, and Keisuke Fujii. 3d pose-based temporal action segmentation for figure skating: A fine-grained and jump procedure-aware annotation approach. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, page 17–26, New York, NY, USA, 2024. Association for Computing Machinery. 1

[40] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[41] Carnegie Mellon University. Cmu graphics lab motion capture database. 3

[42] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 1, 2, 3

[43] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 1

[44] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 4, 5

[45] Calvin Yeung, Kenjiro Ide, and Keisuke Fujii. Autosoccerpose: Automated 3d posture analysis of soccer shot movements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3214–3224, 2024. 5, 7, 8

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 6, 7

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Ic-light github page, 2024. 6, 7

[48] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 2

[49] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image, 2024. 4