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

Jerrin Bright[1,3], Zhibo Wang[1,2,3], Daria Klepachevskyi[1,2,3], Yuhao Chen[1,3], Sirisha Rambhatla[2,3], David Clausi[1,3], John Zelek[1,3]
[1] Vision and Image Processing Lab [2] Critical ML Lab [3] University of Waterloo, Canada

Research Question

Can a controllable synthetic-data pipeline replace real, manually labeled training data in domain-specific pose estimation?

- General-purpose pose datasets cover everyday motion but miss sports, dance, and rehabilitation poses.
- Manual annotation is slow, biased, and not scalable across domains.
- We need synthetic humans with controllable motion, appearance, viewpoint, and scene context.
- Real test: can models trained only on synthetic data transfer to real broadcast footage?

Avatar4D answers this with a single image-to-4D pipeline plus a benchmark across two sports.

Pipeline Inputs

Three dictionaries, no manual labels. Each generated clip samples from three dictionaries to combine appearance, scene, and motion without per-pixel annotation work.

$\mathcal{D}_{\text{img}}$	domain-relevant person images for canonical avatar generation
$\mathcal{D}_{\text{bg}}$	scene / background imagery for in-the-wild composition
$\mathcal{D}_{\text{motion}}$	curated motion sequences from internet sport footage

Motion is grounded in real-world dynamics instead of MoCap-only or text-to-motion diffusion priors.

Syn2Sport Statistics

21k+	4.4M	2
clips	frames	sports

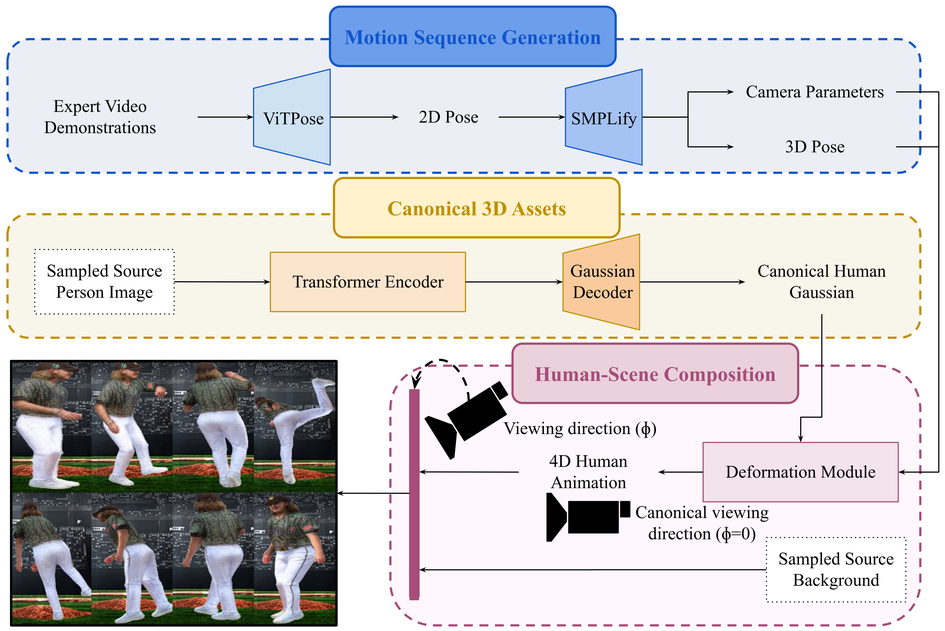
Two sports, fully synthetic. Baseball + ice hockey, train/val splits, no manual labels.

Designed as a substitute for real broadcast training data in domains real datasets do not cover well.

Item	Count
Clips	21,203
Frames	4,483,982
Minutes of play	2,491
Player identities	40

Syn2Sport examples generated by the **Avatar4D** pipeline.

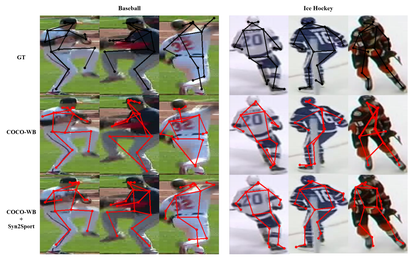
Method



Pipeline. Three decoupled stages generate one labeled synthetic frame.

- **Motion.** Internet sport clips provide realistic pose and camera trajectories.
- **Avatar.** A single source image is lifted into an animatable 3DGS human.
- **Scene.** LBS deforms the avatar, splatting renders RGB/alpha, and backgrounds are sampled per domain.
- **Labels + metadata.** Keypoints are projected per frame, while samples retain camera, motion, identity, background, and dictionary provenance.
- **Why it works.** Target-domain motion preserves sport-specific dynamics, single-image avatars avoid prompt artifacts, and known render state provides supervision automatically.

Zero-Shot Transfer



Synthetic data improves real broadcast pose estimation. TokenPose is trained on COCO-WB, **Syn2Sport**, or both, then evaluated on real sports footage.

Training data	Baseball AP^5	Hockey AP^5
COCO-WB	29.3	13.1
Syn2Sport	23.0	19.5
COCO-WB+**Syn2Sport**	**45.7**	**28.1**

Adding synthetic fine-tuning improves AP^5 by 55.96% on baseball and 114.5% on ice hockey over COCO-WB alone.

Background Ablation

Static scenes transfer best. AP^5 on real transfer; static backgrounds preserve avatar fidelity.

Background	Baseball	Hockey
IC-Light	3.4	1.3
ControlNet	38.1	18.4
Static	**45.7**	**21.7**

References

[1] Qiu et al. "LHM: Large Animatable Human Reconstruction Model from a Single Image." *arXiv*, 2025.

[2] Kerbl et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." *TOG*, 2023.

[3] Black et al. "BEDLAM: Bodies Exhibiting Detailed Lifelike Animated Motion." *CVPR*, 2023.

Headline Result

Synthetic data substitutes for real domain data. Training on **Syn2Sport** alone yields large gains over COCO-WB on real broadcast footage, especially in domains where real data is scarce.

+55.96%	+114.5%	0
AP^5 on real baseball	AP^5 on ice hockey	manual labels used

Best results come from pretraining on COCO-WB and fine-tuning on synthetic samples; ice hockey transfers fail from COCO-WB alone because real datasets lack skating-style motion.

Synthetic Benchmark

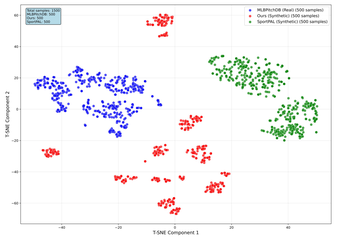
2D pose accuracy on Syn2Sport (AP^5).

Method	Baseball	Ice hockey
DETR	74.5	31.7
HRNet	83.6	43.0
UDP	85.2	38.9
ViTPose	83.6	**55.3**
TokenPose	**89.1**	54.2

TokenPose tops baseball, ViTPose tops ice hockey. Both far exceed DETR/HRNet baselines, showing **Syn2Sport** is informative for current 2D HPE backbones.

Stronger on	static, well-lit pitching deliveries with clear arm trajectories
Weaker on	ice hockey with rapid camera pans, stick occlusion, and skating-only poses
Implication	gap is driven by real-domain noise, not synthetic data quality

Feature-Space Alignment



Synthetic **Syn2Sport** features interleave with real broadcast samples rather than forming a detached synthetic-only cluster.

Acknowledgements

VIP lab
Critical ML
MITACS