Yuxin Yao

yy561@cam.ac.uk +44 7784316157 yuxinyao620.github.io/

Research Profile

PhD researcher in 3D computer vision and embodied AI, developing scalable multimodal models that connect language, 3D perception, and kinematic control. Experienced in large-scale training, VLM conditioning, simulation workflows, and production ML, with publications across NeurIPS, CVPR, and SIGGRAPH.

Education

PhD **University of Cambridge**, Engineering 2023 – 2027

- Supervisor: Prof. Joan Lasenby ; Advisor: Prof. Roberto Cipolla
- **College Senior Scholarship**, Fitzwilliam College, University of Cambridge.

M.Eng. **University College London**, Mathematical Computation 2019 – 2023

- First Class Honours
- Two Research Internships in Computer Vision and Machine Learning
- Research Internship at ETH Zurich (Summer Student Research Fellowship)

Work Experience

Applied Intuition Nov 2026 – Mar 2027
Incoming Machine Learning Research Intern Sunnyvale, CA, USA

Epic Games May 2025 – Dec 2025
Machine Learning Engineer London, UK

- **IP Infringement Detection**: Deployed a partner-facing detection system benchmarked on more than 100,000 images, improving accuracy by over 40% and reducing inference time from 13 hours to 30 minutes through vLLM and distributed multiprocessing.
- **Automated Detection Pipeline**: Built and productionized a scalable YOLO-World pipeline for detecting potential Epic IP infringement across large-scale image datasets.
- **VLM Fine-Tuning for Procedural Content Generation**: Designed data curation and Gemini 2.5 fine-tuning pipelines for VLM-driven scene generation in Unreal Engine's Procedural Content Generation ecosystem.

Research Experience

Instruct-Particulate: Multimodal 3D Articulation and Kinematic Control at Scale

- **Language-Guided 3D Control**: Conditioned a multi-task transformer on part descriptions, connectivity, joint types, and optional 3D point prompts to follow structural constraints and control segmentation and joint prediction.
- **Large-Scale Data Preprocessing**: Scaled the training data to **40x the previously available 3D articulation data** through asset curation, programmatic generation, and VLM pseudo-labeling; built USD/URDF processing pipelines with Blender, DINOv3, and Isaac Sim.
- **Large-Scale Model Training**: Trained the multi-task transformer for 100k steps with a global batch size of 128 across 16 NVIDIA Blackwell GPUs, jointly optimizing part segmentation and joint-motion prediction.
- **Reasoning to Physical Structure**: Combined image-to-3D generation with VLM-generated kinematic prompts to produce simulation-ready articulated assets; introduced the **220-asset Lightwheel benchmark** and achieved nearly **4x the gIoU** of prior baselines.

Particulate: Feed-Forward 3D Part Segmentation and Articulation

- **3D Part Segmentation**: Designed an end-to-end transformer with learned part queries and 3D point features to segment static meshes into functional articulated parts.

- **Physical Structure Prediction**: Jointly predicted kinematic trees, joint types, axes, and motion limits, connecting 3D perception with articulation reasoning and physically meaningful constraints.
- **Fast Inference**: Reconstructed multi-joint articulated structures in seconds, generalized to unseen and AI-generated assets, and exported URDFs for physics simulation.

Gaussian Splatting-Based Camera Pose Estimation

- **Camera Pose Estimation**: Developed efficient batched pose estimation with Gaussian splatting, achieving state-of-the-art performance on small-baseline videos containing moving objects.
- **Foundation Models**: Combined DINOv2 and AnyDepth for feature extraction, depth estimation, and semantic segmentation.
- **Dynamic Motion Modeling**: Modeled non-rigid object motion with a PyTorch neural motion field regularized by semantic video priors.

AutoKeyframe: Controllable Human Motion Synthesis and Editing

- **Action-Conditioned Motion**: Developed a transformer-based autoregressive diffusion model conditioned on 3D root trajectories and action labels, with skeleton-guided constraints for controllable generation and localized editing.
- **Results**: Reduced FID by **29% for motion generation and 72% for motion editing**, while mitigating foot skate and ground penetration.

Publications and Talks

[1] Ruining Li*, Yuxin Yao*, Matt Zhou, Chuanxia Zheng, Christian Rupprecht, Joan Lasenby, Shangzhe Wu, Andrea Vedaldi. "Instruct-Particulate: Scaling Feed-Forward 3D Object Articulation with Kinematic Control". In: *Advances in Neural Information Processing Systems (NeurIPS) 2026*.

[2] Zhening Huang*, Yueyan Li*, Johnathan Chiu*, Xiaoyang Lyu, Matt Zhou, Yuxin Yao, Joan Lasenby, Shangzhe Wu. "LiteReality-Agent: An Agentic System for Interactable 3D Indoor Scene Reconstruction". *arXiv:2610.01863*, 2026.

[3] Ruining Li*, Yuxin Yao*, Chuanxia Zheng, Christian Rupprecht, Joan Lasenby, Shangzhe Wu, Andrea Vedaldi. "Particulate: Feed-Forward 3D Object Articulation". In: *(CVPR) 2026*.

[4] Yuxin Yao, Yan Zhang, Zhening Huang, Joan Lasenby "SmallGS: Gaussian Splatting Based Camera Pose Estimation for Small-Baseline Videos". In: *(CVPR Proceedings) 2025*

[5] Bowen Zheng, Ke Chen, Yuxin Yao, Zijiao Zeng, Xinwei Jiang, He Wang, Joan Lasenby, and Xiaogang Jin. "AutoKeyframe: Autoregressive Keyframe Generation for Human Motion Synthesis and Editing". In: *(SIGGRAPH Conference Papers '25)*.

[6] Yuxin Yao, Christian Hockey, Joan Lasenby. "Simplifying and Generalising Equivariant Geometric Algebra Networks". In: *AGACSE*. 2024.

[7] Yao Yan, Thomas Schaffter, Timothy Bergquist, ... Yuxin, Yao, ... Sean Mooney. "A Continuously Benchmarked and Crowdsourced Challenge for Rapid Development and Evaluation of Models to Predict COVID-19 Diagnosis and Hospitalization". In: *JAMA Network Open*, 4(10), e2124946. 2021.

[8] **Contributed Talk**: "Simplifying and Generalising Equivariant Geometric Algebra Networks". In: *AGACSE* (Amsterdam, NL). 2024.

[9] **Invited Talk**: Feed-forward 3D object articulation and kinematic control. In: *ASPIRE Computer Vision Workshop*, Oxford VGG, University of Oxford (Oxford, UK). 2026.

[10] **Invited Talk**: Feed-forward 3D object articulation and kinematic control. In: *Spatial Intelligence Seminar*, hosted by Junwei Zhang's group at the Technical University of Munich (Munich, Germany). 2026.

Skills and Interests

Technical: Python (7 years), PyTorch (6 years), Lightning, Hugging Face, scikit-learn, and vLLM; distributed training and inference, scalable research pipelines, simulation workflows, and production deployment.
Research Methods: Multimodal foundation models, VLM conditioning and fine-tuning, transformers, diffusion models, 3D and temporal representation learning, embodied AI, and generative modeling.