

Zinuo (Henry) You

GENERATIVE VIDEO & WORLD MODELS | LOW-NFE INFERENCE & DISTILLATION | CONTINUOUS-TIME GENERATIVE MODELS

EMAIL
zinuo.you@bristol.ac.uk

GITHUB
github.com/pixelhero98

GOOGLE SCHOLAR
Google Scholar profile

PROFILE

PhD candidate in Computer Science at the University of Bristol, leading modelling research and implementation across three UKRI-funded projects. Work spans long-horizon T2V/Action2V, representation/KV drift, LoRA post-training, and low-step inference/distillation for flow matching and neural ODEs. First-author research includes an ICML 2026 Spotlight and publications at AAAI and ICASSP.

EDUCATION

University of Bristol – PhD in Computer Science	2022 – Present
University of Sheffield – MSc in Electronic and Electrical Engineering, Distinction	2019 – 2020
Southwest University – Bachelor’s in Electronic Science and Technology and Solid-State Physics	2015 – 2019

TECHNICAL SKILLS

Video and World Models	Transformers/DiT/MMDiT, T2V/I2V/Action2V, causal/autoregressive video, temporal consistency, KV/hidden-state drift, camera/action conditioning.
Generative and Efficient Inference	VAE/latent diffusion, flow matching, neural ODEs/SDEs, low-NFE sampling, schedule/time reparameterisation, consistency/flow-map distillation.
Training, Evaluation, and Deployment	Python/PyTorch, LoRA/PEFT, SFT/DPO, multi-GPU/H100, vLLM, INT8/NPU, Slurm/HPC; generation quality, inference efficiency, temporal consistency, and OOD/robustness evaluation.

GENERATIVE-MODEL RESEARCH AND PROJECTS

Long-Horizon Consistency for Causal Video & Interactive World Models Mar 2026 – Present

UKRI / VIDEO GENERATION CONSISTENCY

- Led the Wan2.1-1.3B + Causal-rCM long-horizon T2V/Action2V evaluation pipeline, analysing autoregressive error accumulation over $5 \rightarrow 10 \rightarrow 20 \rightarrow 40+$ chunks; covered prompt/motion difficulty, action reversals, and repeated/unseen controls, with VBench-Long/I2V and WBench/WorldRoamBench for consistency and action adherence.
- Designed probing of VAE latents, DiT/MMDiT hidden states, and KV histories; compared online KV with refreshed or recomputed histories and validated HunyuanVideo-1.5-8B + minWM and Wan2.2-5B + BiWM. Short-horizon drift predicted long-horizon drift at Spearman $\rho = 0.61\text{--}0.70$, approximately 15%–20% above direction-only/latent-norm baselines.
- On $2 \times \text{H100}$, ran 500–1,000-step partition-consistency LoRA: reduced rollout inconsistency by approximately 25% and long-horizon drift by approximately 20%, kept short-horizon quality within approximately 2%, and improved VBench-Long/WBench consistency/action metrics by 3%–8%.

Low-NFE Inference Path Optimisation for Generative ODEs 2024 – Present

WORKING PAPER / CODE

- Relaxed discrete inference grids into normalised time densities, learning a context-conditioned low-step path for a frozen vector field, ODE solver, and target NFE; consistency/flow-map distillation maps the optimised path to one-step generation of comparable quality.
- Designed a teacher–student optimiser: the teacher learns schedule utility via pairwise ranking and Huber regression; mixture targets, forward KL, and downstream rewards train a student conditioned on context, solver, and NFE.
- Across nine public benchmarks and NFE 4/8/10/12/16, improved core metrics by 4%–22% over uniform, adaptive, and heuristic schedules and generalised zero-shot to unseen NFEs; on ImageNet-64/256, FID improved by 5%–12%, with Precision/Recall stable or up 1–3 points.

Latent Laplace Diffusion for Irregular Temporal Sequences 2023 – Present

ICML 2026 SPOTLIGHT

- Developed LLapDiff, learning conditional diffusion in VAE latent space and generating forecast windows in parallel at query timestamps, avoiding regriding distortion and stepwise integration.
- Parameterised stable latent trajectories with learnable complex-conjugate poles and damping constraints, combining renewal averaging with interval-aware history encoding for irregular sampling.
- ICML 2026 Spotlight (top 2.2%): across seven datasets covering extrapolation, interpolation, and OOD shift, achieved $1.2 \times \text{--} 130 \times$ faster inference and 11%–29% lower CRPS/MSE/MAE under matched protocols, including 2%–92% missingness.

GENERATIVE-MODEL RESEARCH AND PROJECTS

Refractive Clocks: Diffusion-to-Flow Schedule Transfer

2023 – Present

[NEURIPS 2026 UNDER REVIEW](#)

- Introduced the refractive-clock view, interpreting non-uniform schedules as temporal reparameterisations and transferring AYS, GITS, and OTS from diffusion to frozen flow-matching backbones.
- Derived the reparameterised dynamics and material derivative, proving continuous-limit endpoint invariance while finite-step discretisation changes the induced dynamics and inference-time density.
- Proposed local time-density alignment; experiments showed that local integration fidelity and consistency alone do not determine downstream sample quality.
- Built a unified evaluation over five extrapolation datasets, three conditional-generation settings, four fixed solvers, and NFE 4/8/10/12/16; GITS improved core extrapolation metrics by 6.1%–10.4%, with conditional gains dependent on task and metric.

GNN Architecture Capacity and Financial Graph Learning

2023 – 2025

[AAAI 2026 / ICASSP 2024](#)

- Led channel-capacity-constrained GNN architecture search, reducing search time by approximately $10\times$ and improving accuracy/F1 by 2%–11%; also developed dynamic stock-graph models using relational retention, graph diffusion, and decoupled representations, published at AAAI 2026, ICASSP 2024, and ICAART.

OTHER UKRI-FUNDED PROJECTS

Restricted-Domain LLM Audit Agents

Feb 2026 – Present

[UKRI / LLM AGENT](#)

- Led development of an evidence-constrained agent for 20 audit-task categories over millions of UK corporate filings from the previous five years; created approximately 20,000 SFT examples and 20,000 preference pairs, and ran parameter-efficient SFT/DPO on 30B-class open models using up to $2\times H100$.
- Built programmatic evaluators and a vLLM batched-assessment pipeline; achieved 81% task pass, 79% correct refusal, and 76% correct evidence citation, outperforming zero-shot and RAG-CoT agent baselines.

TPE-AS: Adaptive Bayesian Optimisation for Black-Box Portfolio Models

Jun 2024 – Apr 2025

[UKRI / STRATIPHY / ICAAI 2025](#)

- Led the formulation of a partner's private portfolio-model tuning as a black-box problem with at least 25 mixed variables, expensive noisy evaluations, and a limited budget, interacting through a controlled evaluator isolated from private models, market data, and backtesting infrastructure.
- Proposed TPE-AS, adding a budget-dependent mean–variance Lagrangian penalty and clipped importance correction to good/bad Parzen-density search for improved stability and sample efficiency.
- Across three black-box models, four backtesting settings, and 500 evaluations per scenario, achieved the highest annualised Sharpe in 8/12 scenarios and the lowest optimisation-trajectory variance in 12/12; the public repository reproduces the optimiser and transparent synthetic experiments.

INDUSTRY EXPERIENCE

Automotive Engineer Intern

Apr 2021 – Nov 2021

[QUALCOMM](#)

- Evaluated and integrated real-time automotive object/trajectory detection; using FP32 as the accuracy anchor, supported INT8 output alignment, operator compatibility, and deployment diagnosis on the target NPU.
- Reached approximately 40 FPS in steady-state device tests and improved benchmarks and failure attribution across quantisation accuracy, module outputs, and runtime performance.

PUBLICATIONS AND MANUSCRIPTS

- [1] You, Z., Zheng, J., & Cartlidge, J. (2026). Latent Laplace Diffusion for Irregular Multivariate Time Series. ICML 2026. Spotlight.
- [2] You, Z., Zheng, J., & Cartlidge, J. (2026). Refractive Clocks: Diffusion-to-Flow Schedule Transfer for Conditional Time Series Flow Matching. NeurIPS 2026, under review.
- [3] You, Z., Zheng, J., & Cartlidge, J. (2026). How Wide and How Deep? Mitigating Over-squashing of GNNs via Channel Capacity Constrained Estimation. AAAI 2026.
- [4] You, Z., Cartlidge, J., Elliott, K., Ge, M., & Gold, D. (2025). Improving Bayesian Optimization for Portfolio Management with Adaptive Scheduling. ICAAI 2025.
- [5] You, Z., Zhang, P., Zheng, J., & Cartlidge, J. (2024). Multi-Relational Graph Diffusion Neural Network with Parallel Retention for Stock Trends Classification. ICASSP 2024.
- [6] You, Z., Shi, Z., Bo, H., Cartlidge, J. P., Zhang, L., & Ge, Y. (2024). DGDNN: Decoupled Graph Diffusion Neural Network for Stock Movement Prediction. ICAART 2024.
- [7] You, Z., Shi, Z., Bo, H., Cartlidge, J., Zhang, L., & Ge, Y. (2025). Multi-Stock Movements Classification: A Decoupled Graph Diffusion Neural Network Approach. LNCS / ICAART 2025.

HONOURS AND FUNDING

University of Bristol Scholarship Doctoral scholarship / 2025

Jean Golding Institute Seed Corn Fund Seed funding / 2023

ECEF Fellowship Award Fellowship / 2024

MCM/ICM Honorable Mention Mathematical modelling / 2019