Junjie Qiu

INFORMATION FOR INTELLIGENT ACTION

junjieqiu.com jay.junjie.qiu@outlook.com (+86) 191-2013-2361

01 Personal Profile

I am a first-year Ph.D. student at HKUST (GZ), interested in **information for intelligent action**: how intelligent systems learn, retain, and use information to understand changing environments and guide action. My work spans learning temporal representations from history, maintaining narrative context, and using feedback to guide video generation. I co-authored a NeurIPS 2025 Spotlight paper. Previously at LuckyShort, I was an Architect reporting directly to the CTO. I also worked at Huawei's 2012 Laboratories and SUSTech's HPC center.

02 Education

Hong Kong University of Science and Technology (GZ) Sep. 2026 - Present

Ph.D. in Intelligent Transportation

Southern University of Science and Technology Sep. 2021 - Jul. 2025

B.S. in Data Science and Big Data Technology

GPA: 3.89/4.00 · Rank: 4/53 (top 10%)

Honors: Outstanding Graduate; Outstanding Undergraduate Thesis;
Top Ten Graduate Award, Shude Residential College (Jun. 2025)

03 Industry Experience

ByFusion · LuckyShort Mar. 2026 - Sep. 2026

Architect · AI short-drama production

- Led video-agent architecture for long-form storytelling, coordinating narrative planning, visual reference assets, controllable generation, and multimodal evaluation across an end-to-end production workflow.
- Designed application-level memory to retain character and narrative state, check new events against prior context, and update downstream context after revisions.
- Integrated agent-driven script production, video generation, publishing, and ad delivery for a platform reaching 3B+ TikTok/YouTube views, including multiple short-drama videos with 100M+ views each.
- Linked generated videos with their context, user selections, and rejection reasons, preserving feedback for visual-quality evaluation and cross-shot consistency checks.

Huawei · 2012 Laboratories Mar. 2025 - Dec. 2025

Research Intern

- Adapted pretrained temporal models to Huawei Ascend, preparing telecom and power time series for model training and evaluation across heterogeneous industrial forecasting datasets.
- Reproduced forecasting baselines and evaluated temporal models on Huawei's internal telecom and power datasets, validating the strongest forecasting results among the methods tested.
- Contributed to FactoST [1]; the paper reports MAE reductions of 12.5% on Electricity vs. OpenCity and 13.7% on Weather vs. Moirai, with 10% training data and 96-step context and forecast horizons.

SUSTech · Center for Computational Science and Engineering Mar. 2024 - Jul. 2024

Intern · High-Performance Computing

- Conducted HPL (LINPACK) and HPCG benchmarking on the Qiming and Taiyi supercomputers, establishing performance baselines for computational throughput and distributed execution efficiency.
- Provisioned and commissioned HPC compute nodes, covering hardware assembly, software deployment, and system debugging to integrate new computing capacity into existing supercomputing clusters.
- Profiled MPI communication to identify bottlenecks, then optimized switch interconnect topology for four nodes installed in Jun. 2024, reducing measured MPI communication time by approximately 20%.

04 Research Experience

FactoST-v2 · Universal Temporal Pretraining Oct. 2025 - Jan. 2026

UTP method design and experiments · Preprint [2]

- Designed and pretrained an encoder-only quantile model to learn from normalized sequence patches and masked history prefixes, supporting up to 2,048 input and 256 forecast steps.
- Evaluated fixed-weight zero-shot forecasts on benchmarks excluded from pretraining ($12 \to 12$, $96 \to 96$); masking reduced PEMS08 MAE by 15% vs. no masking (107.22 to 91.09 at $96 \to 96$; Tables IX-X in [2]).

FactoST · Cross-domain Temporal Pretraining Feb. 2025 - May 2025

Cross-domain transfer evaluation · NeurIPS Spotlight [1]

- Evaluated zero-shot and few-shot transfer across diverse forecasting datasets, assessing cross-domain generalization and data-efficient adaptation through the reuse of pretrained temporal representations.

05 Technical Projects

Agentic Video Generation · Cross-Shot Consistency Jul. 2026 - Sep. 2026

ByFusion · LuckyShort · Internal project

- Built reference-conditioned keyframe generation with Gemini 3.8 Flash checks for identity and scene drift, routing feedback to regenerate faulty shots while retaining accepted frames and shared visual references.
- Centralized creative direction across generation stages and refined it with final-video feedback, warm-starting candidate search from model priors to guide exploration of narrative and visual alternatives.

Rare-Earth Pricing · Action-Conditioned Replanning Jan. 2026 - Mar. 2026

China Mobile Research Institute · China Rare Earth Group

- Contributed to long-horizon price modeling and experiments, conditioning forecasts on geopolitical and supply-chain observations and pricing actions to compare market scenarios and support replanning.

Tencent Cloud · Behavior-Aware Resource Allocation May 2025 - Aug. 2025

Research collaboration · Tencent Rhino-Bird Fund

- Contributed to temporal modeling for Tencent Cloud resource allocation, using behavioral profiles and historical allocation sequences as decision context for user-specific compute resource assignment.

ASC24 · LLM Inference Infrastructure & Optimization Apr. 2024

Ranked 10th among 300+ teams · First Prize & Group Competition Award

- Built distributed LLM serving on four nodes with two GPUs per node, integrating vLLM inference engines, Ray orchestration, and RDMA networking for concurrent execution under competition constraints.
- Implemented dynamic request scheduling across Ray-managed vLLM replicas, dispatching tasks to idle instances through an available-worker queue to balance workloads on a fixed GPU cluster.

06 Skills

Physical AI	Concepts: Diffusion / Rectified Flow, Multimodal Conditioning, DreamerV3, JEPA, RSSM, Latent Dynamics, Latent Imagination, Model-based RL, Action Abstraction, Steerable VLA Policies.
Memory & Context	Concepts: Belief-State Modeling, Multimodal Memory, Event-centric Memory, Prospective Memory.
Training, Inference & Simulation	Project work: vLLM, Ray, RDMA, Huawei Ascend. Familiarity: 4D (DP/TP/PP/CP), Continuous Batching, MuJoCo.

07 Bibliography

[1] S. Zhong, J. Qiu, Y. Wu, X. Zou, Z. Rao, B. Yang, C. Guo, H. Xu, and Y. Liang, "Learning to Factorize Spatio-Temporal Foundation Models," in Advances in Neural Information Processing Systems (NeurIPS), 2025. **Spotlight**.

[2] S. Zhong, J. Qiu, Y. Wu, Y. Liu, Y. He, Z. Rao, B. Yang, C. Guo, H. Xu, and Y. Liang, "Learning to Factorize and Adapt: A Versatile Approach Toward Universal Spatio-Temporal Foundation Models." arXiv preprint arXiv:2601.12083, 2026.