# Research Statement

Joey Tianyi Zhou
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6828 0568; Email: tyzhou@smu.edu.sg
18 May 2026

## Background

Modern AI has advanced by scaling data, models, and computation, but the same scaling has exposed practical limits. Training and deployment remain expensive; predictions can shift under perturbations, bias, adversarial inputs, or distribution change; and foundation models still struggle when reasoning must be grounded in evidence and sustained over long decision processes. My research focuses on **building AI systems that learn efficiently, behave reliably, and support accountable decision-making in real-world environments**.

The central question is how AI systems can learn more from less data, remain dependable outside clean benchmarks, and use foundation models as reasoning and action systems. My work starts from the view that data shape more than accuracy. They affect which capabilities a model acquires, which shortcuts it exploits, which biases it inherits, and which vulnerabilities enter the learning pipeline. This view connects three research areas: Resource-efficient AI, Safe and Reliable AI, and Large Foundation Models and Agentic Frameworks.

My long-term goal is to develop **powerful, efficient and trustworthy agentic AI systems** that use data and computation deliberately, reason with multimodal evidence, act under real constraints and remain auditable when their decisions matter. This agenda is consistent with Singapore's need for AI methods that can be deployed in advanced manufacturing, logistics, finance, healthcare and public-facing services under constraints on data, computation, privacy and regulation.

## Research Areas

I organize this agenda into three connected research areas. The first develops methods that reduce the data, compute, memory, and energy required for learning and deployment. The second studies reliability, robustness, bias, and security risks that emerge across the learning pipeline. The third integrates these foundations into large foundation models and agentic frameworks that can reason, use tools, interact with environments, and support accountable decisions.

## Resource-efficient AI

My first line of work studies the resource cost of machine intelligence. A basic question in machine learning is sample efficiency: how many examples are needed for a model to generalize? In modern AI, this question has widened to include annotation cost, training energy, memory footprint, inference latency, and deployment constraints. Edge devices cannot host large models without compression. Medical and scientific

datasets are expensive to annotate. Multimodal datasets are costly to align. Large language model training is constrained by model size, data quality, redundancy, privacy, and energy cost. These constraints lead to a concrete question: what information is truly necessary for learning, and how can we preserve it with much less data and computation?

My recent work approaches this question by treating datasets as objects to be optimized. Dataset condensation synthesizes compact training signals that preserve the learning effect of the original data. Across this line of work, I study what a small dataset must retain in order to train a model well: training dynamics, class structure, diversity, generalization signals, and flexibility across compute budgets [4, 5, 7, 8]. Data efficiency depends on identifying which learning signals must survive compression.

This work has recently expanded from standard vision tasks to multimodal and language-model settings. In these regimes, the value of data lies less in isolated examples than in the relations and behaviors they induce: cross-modal alignment, reward-guided behavior, instruction following, semantic coverage, and downstream reasoning [1, 2]. This shift moves dataset condensation closer to foundation model training, where compact data must preserve capabilities rather than labels alone.

Beyond data condensation, I have also studied resource-efficient learning under hardware, optimization, and transfer constraints. This includes reusing knowledge when labeled data are scarce, improving training efficiency, and adapting models to limits in precision, memory, and energy [16-23]. The area now links data, optimization, transfer, and deployment as parts of the same efficiency problem.

I will extend this line of work to data condensation and pruning for foundation model training, including instruction data, multimodal corpora, medical data and domain specific knowledge. The key difficulty is that base models learn many capabilities at once, so compact data must preserve not only accuracy, but also reasoning, calibration, fairness, robustness and safety. I will also study resource-aware training and deployment where computation, latency, memory, communication, privacy, and energy are optimized together. A condensed dataset or compressed model that preserves benchmark accuracy while amplifying bias, backdoors, or overconfidence does not solve the real deployment problem. This work is important for sustainable AI, edge intelligence, personalized AI and national AI adoption, where practical systems regularly operate under tight data, latency, privacy and infrastructure constraints.

## Safe and Reliable AI

My second line of work studies the gap between model performance and model reliability. Test accuracy captures one form of generalization; deployment demands a stronger one. Models need to remain reliable when the data distribution changes, when evidence is incomplete, when modalities disagree, when inputs are adversarial, and when social or security risks are present. A model can achieve high benchmark accuracy while relying on spurious correlations, producing overconfident predictions from weak evidence, encoding social bias, or carrying hidden backdoors introduced

through training data, prompts, or compressed representations. I study how these failure modes arise, how they can be exposed, and how they can be controlled.

One part of my work studies adversarial robustness, backdoor attacks, and data poisoning. I view attacks as diagnostic tools: they reveal which assumptions a model silently depends on. Across language, black-box, and 3D vision settings, this work separates clean accuracy from robustness under adaptive and domain-specific threats [14, 15, 24].

More recently, my work examined how vulnerabilities shift with new learning paradigms, including prompt-based learning, dataset distillation and large multimodal models [6, 25, 26]. Safety problems migrate across these cases into the least inspected part of an AI workflow: prompts, compressed data, or the correspondence between modalities.

Another strand examines reliability, uncertainty and explainability. In multimodal and multiview learning, different sources of evidence may disagree, and their trustworthiness can vary across samples and contexts. My work develops methods that estimate the evidence quality, convey uncertainty and provide explanations when labels, modalities or predictions are incomplete or unreliable [10-13]. Recently, I have studied social bias knowledge within language models, focusing on how biased associations can be identified and mitigated rather than treated as incidental artifacts [9].

I therefore treat safety as a property of the full learning pipeline, not only of the final model. A backdoor in a distilled dataset, an unreliable modality in a fusion model, a biased association in a language model, or an overconfident agentic decision can each become a system-level failure.

I will next study safety in synthetic, compressed and privacy-preserving data regimes where raw data may be inaccessible and traditional inspection methods may fail. I will also study multimodal safety, particularly attacks that break the correspondence between language, vision, video and structured signals. A further direction is dependability in agentic systems where failures can accumulate as agents retrieve information, use tools, interact with users or coordinate with other agents. In these settings, reliability, bias control, safety testing and auditability must be built into the system design from the start.

## Large Foundation Models and Agentic Frameworks

My third line of work brings the previous two together. Efficient learning makes foundation models cheaper to adapt and deploy; reliable learning makes their behavior testable when they leave clean benchmarks. Building on these foundations, I study how large models can be used as reasoning and action systems. The technical challenge is no longer only to produce a correct answer in one pass. A useful AI system often has to define a task, gather evidence, choose tools, execute actions, observe the result, and revise its plan.

My recent work approaches this problem through agentic frameworks. A common way to build agents is to design a top-down workflow: decompose a task, call tools, reflect,

and repeat. Such workflows support task decomposition and tool use, while their manual design often limits adaptation across tasks. I am interested in a complementary view in which agents acquire and refine reusable skills through repeated interaction [27]. I also study numerical and quantitative reasoning in market-like and digital-twin environments, where reasoning errors have concrete consequences and feedback can be folded back into the agent's behavior [28, 29].

These loops need perceptual anchors. My work on spatio-temporal grounding, sensory action understanding, visual grounding, video localization, crowd analysis, anomaly detection, and medical vision-language knowledge mining connects high-level language and planning to visual, temporal, medical, and environmental signals [30-36]. Without such feedback, an agent's plan remains detached from the conditions it is supposed to affect.

This work pushes evaluation beyond fluency and one-shot task success. For foundation models and agents, we need to examine how they use observations, revise behavior, manage uncertainty, and expose actions for audit over long horizons. The three areas meet at this point: Resource-efficient AI reduces the cost of repeated interaction; Safe and Reliable AI limits failures as decisions accumulate; agentic AI organizes those interactions so that systems can improve over time.

My next goal is to build agentic foundation model systems with three capabilities: grounding in multimodal evidence, improvement through interaction and feedback, and explicit reporting of uncertainty, evidence, and decision processes. As large models become interfaces for work, education, science, finance, health, and public information, users will need tools that augment judgment and productivity while remaining governable and adaptable to local needs.

**Selected Publications and Outputs**

1. Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and **Joey Tianyi Zhou**. Beyond Modality Collapse: Representation Blending for Multimodal Dataset Distillation. NeurIPS 2025.

2. Cheng Shen, Yew-Soon Ong, and **Joey Tianyi Zhou**. CondenseLM: LLMs-driven Text Dataset Condensation via Reward Matching. EMNLP 2025.

3. Xin Zhang, Jiawei Du, Ping Liu, and **Joey Tianyi Zhou**. Breaking Class Barriers: Efficient Dataset Distillation via Inter-Class Feature Compensator. ICLR 2025.

4. Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and **Joey Tianyi Zhou**. Diversity-Driven Synthesis: Enhancing Dataset Distillation through Directed Weight Adjustment. NeurIPS 2024 Spotlight.

5. Yang He, Lingao Xiao, **Joey Tianyi Zhou**, and Ivor Tsang. Multisize Dataset Condensation. ICLR 2024 Oral.

6. Xiaopeng Xie, Ming Yan, Xiwen Zhou, Chenlong Zhao, Suli Wang, Yong Zhang, and **Joey Tianyi Zhou**. Shortcuts Arising from Contrast: Towards Effective and Lightweight Clean-Label Attacks in Prompt-Based Learning. EMNLP 2024.

7. Yang He, Lingao Xiao, and **Joey Tianyi Zhou**. You Only Condense Once: Two Rules for Pruning Condensed Datasets. NeurIPS 2023.

8. Jiawei Du, Qin Shi, and **Joey Tianyi Zhou**. Sequential Subset Matching for Dataset Distillation. NeurIPS 2023.

9. Ruizhe Chen, Yichen Li, Jianfei Yang, Yang Feng, **Joey Tianyi Zhou**, Jian Wu, and Zuozhu Liu. Identifying and Mitigating Social Bias Knowledge in Language Models. NAACL 2025.

10. Zongbo Han, Changqing Zhang, Huazhu Fu, and **Joey Tianyi Zhou**. Trusted Multi-View Classification. ICLR 2021.

11. Zongbo Han, Changqing Zhang, Huazhu Fu, and **Joey Tianyi Zhou**. Trusted Multi-View Classification with Dynamic Evidential Fusion. IEEE TPAMI 2022.

12. Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, **Joey Tianyi Zhou**, and Qinghua Hu. Trustworthy Multimodal Regression with Mixture of Normal-inverse Gamma Distributions. NeurIPS 2021.

13. Xi Peng, Yunfan Li, Ivor Tsang, Hongyuan Zhu, Jiancheng Lv, and **Joey Tianyi Zhou**. XAI Beyond Classification: Interpretable Neural Clustering. JMLR 2022.

14. Xinke Li, Zhiru Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and **Joey Tianyi Zhou**. PointBA: Towards Backdoor Attacks in 3D Point Cloud. ICCV 2021.

15. Di Jin, Zhijing Jin, **Joey Tianyi Zhou**, and Peter Szolovits. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. AAAI 2020.

16. Ming Yan, Hao Zhang, Di Jin, and **Joey Tianyi Zhou**. Multi-source Meta Transfer for Low Resource Multiple Choice Question Answering. ACL 2020.

17. **Joey Tianyi Zhou**, Hao Zhang, Di Jin, and Xi Peng. Dual Adversarial Neural Transfer for Sequence Labeling. IEEE TPAMI 2021.

18. **Joey Tianyi Zhou**, Sinno Jialin Pan, and Ivor W. Tsang. A Deep Learning Framework for Hybrid Heterogeneous Transfer Learning. Artificial Intelligence 2019.

19. **Joey Tianyi Zhou**, Ivor W. Tsang, Sinno Jialin Pan, and Mingkui Tan. Multi-class Heterogeneous Domain Adaptation. JMLR 2019.

20. Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Y. F. Tan, and **Joey Tianyi Zhou**. Sharpness-aware Training for Free. NeurIPS 2022.

21. Jiawei Du, Hanshu Yan, Jiashi Feng, **Joey Tianyi Zhou**, Liangli Zhen, Rick Siow Mong Goh, and Vincent Y. F. Tan. Efficient Sharpness-aware Minimization for Improved Training of Neural Networks. ICLR 2022.

22. Zhehui Wang, Tao Luo, Rick Siow Mong Goh, and **Joey Tianyi Zhou**. EDCompress: Energy-Aware Model Compression for Dataflows. IEEE TNNLS 2022.

23. Tian Huang, Tao Luo, and **Joey Tianyi Zhou**. Adaptive Precision Training for Resource Constrained Devices. ICDCS 2020.

24. Jiawei Du, Hu Zhang, **Joey Tianyi Zhou**, Yi Yang, and Jiashi Feng. Query-efficient Meta Attack to Deep Neural Networks. ICLR 2020.

25. Ziyuan Yang, Ming Yan, Yi Zhang, and **Joey Tianyi Zhou**. Poisoned Distillation: Injecting Backdoors into Distilled Datasets Without Raw Data Access. AAAI 2026.

26. Zhifang Zhang, Qiqi Tao, Jiaqi Lv, Na Zhao, Lei Feng, and **Joey Tianyi Zhou**. TokenSwap: Backdoor Attack on the Compositional Understanding of Large Vision-Language Models. ICML 2026.

27. Jiawei Du, Jinlong Wu, Chen Yuzheng, Yucheng Hu, Bing Li, and **Joey Tianyi Zhou**. Rethinking Agent Design: From Top-Down Workflows to Bottom-Up Skill Evolution. arXiv.

28. Tianmi Ma, Jiawei Du, Wenxin Huang, Wenjie Wang, Liang Xie, Xian Zhong, and **Joey Tianyi Zhou**. Agent Trading Arena: A Study on Numerical Understanding in LLM-Based Agents. EMNLP 2025.

29. Tianmi Ma, Wenxin Huang, Jiawei Du, Lin Li, Xian Zhong, and **Joey Tianyi Zhou**. Evolving Quantitative Reasoning through Self-Play in Digital Twin Markets. ICML 2026.

30. Heng Zhao, Yew-Soon Ong, and **Joey Tianyi Zhou**. Agentic Spatio-Temporal Grounding via Collaborative Reasoning. SIGIR 2026.

31. Bing Li, Haotian Duan, Yun Liu, Le Zhang, Wei Cui, and **Joey Tianyi Zhou**. STADe: Sensory Temporal Action Detection via Temporal-Spectral Representation Learning. IEEE TPAMI 2025.

32. Heng Zhao, **Joey Tianyi Zhou**, and Yew-Soon Ong. Word2Pix: Word to Pixel Cross-Attention Transformer in Visual Grounding. IEEE TNNLS 2024.

33. Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, **Joey Tianyi Zhou**, and Rick Siow Mong Goh. Natural Language Video Localization: A Revisit in Span-based Question Answering Framework. IEEE TPAMI 2021.

34. **Joey Tianyi Zhou**, Le Zhang, Jiawei Du, Xi Peng, Zhiwen Fang, Zhe Xiao, and Hongyuan Zhu. Locality-Aware Crowd Counting. IEEE TPAMI 2021.

35. **Joey Tianyi Zhou**, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. AnomalyNet: An Anomaly Detection Network for Video Surveillance. IEEE TIFS 2019.

36. Jiaxiang Liu, Tianxiang Hu, Jiawei Du, Ruiyuan Zhang, **Joey Tianyi Zhou**, and Zuozhu Liu. KPL: Training-Free Medical Knowledge Mining of Vision-Language Models. EMNLP 2024.