Fengfei Yu

• San Diego, California, USA • alvinyu025.github.io • alvinfengfei@gmail.com

EDUCATION

University of California, San Diego — San Diego, California, USA
M.S. in Computer Science and Engineering — Sep 2025 – Present

Hong Kong Baptist University — Kowloon Tong, Kowloon, Hong Kong
B.Sc. in Computer Science, Minor in Applied Mathematics — Sep 2021 – Jun 2025

McGill University — Montreal, Quebec, Canada
Visiting Student, Department of Computer Science — Aug 2024 – Dec 2024

RESEARCH INTERESTS

Trustworthy Machine Learning, Uncertainty Quantification, Reinforcement Learning, Large Language Models.

RESEARCH EXPERIENCE

Rose-STL Lab, UC San Diego — Oct 2025 – Present
Advisors: Rose Yu (Associate Professor) & Yian Ma (Assistant Professor)
· Developed CSR, an RL framework calibrating LLMs in semantic space without a verbalized-confidence interface.
· Designed a semantic calibration reward: agreement among correct rollouts, exploration among wrong ones.
· Reduced ECE by up to 40% and improved AUROC by up to 31% over verbalized-confidence baselines.
· Generalizable calibration across HotpotQA, TriviaQA, MSMARCO, NQ-Open and three model families.

TMLR Group, Hong Kong Baptist University — Jan 2025 – Jun 2025
Advisor: Bo Han (Associate Professor)
· Proposed AlignMI on the manifold hypothesis: priors denoise inversion-loss gradients via the generator manifold.
· Pinpointed gradient–manifold alignment as the dominant predictor of target-model vulnerability to MIAs.
· Designed a training objective enforcing such alignment, yielding a principled MIA-robustness diagnostic.
· Validated on CelebA (LOMMA/DCGAN) and FFHQ (PPA/StyleGAN); AlignMI accepted at NeurIPS'25.

TMLR Group, Hong Kong Baptist University — Jun 2024 – Oct 2024
Advisor: Bo Han (Associate Professor)
· Proposed DDMI by distilling diffusion models into single-step MIA priors, removing GAN optimization instability.
· Exposed privacy vulnerabilities of multimodal models such as CLIP under diffusion-prior MIAs.
· Derived theoretical intuition for diffusion-prior inversion behavior and ran ablations on timestep schedules.
· Achieved higher reconstruction fidelity and more stable inversion gradients than GAN-based MIA baselines.

TMLR Group, Hong Kong Baptist University — Jun 2024 – May 2025
Advisor: Bo Han (Associate Professor)
· Proposed DCC, an OOD-detection defense: classifier OOD weakness leaks signal to generative MIAs.
· Quantified a direct correlation between OOD-detection quality and MIA reconstruction fidelity across attacks.
· Developed a classifier training strategy hardening OOD detection without sacrificing in-distribution accuracy.
· Benchmarked against SOTA generative MIAs with consistent robustness gains at minimal accuracy cost.

TMLR Group, Hong Kong Baptist University — Jan 2024 – May 2024
Advisor: Bo Han (Associate Professor)
· Contributed to PPDG, minimizing the distributional gap between pseudo-private and private training data.
· Fine-tuned the generator on pseudo-private samples to raise the density of true private data under the prior.
· Implemented the end-to-end training pipeline and ran main experiments and ablations in low-resolution settings.
· Benchmarked against SOTA generative MIAs; PPDG accepted at NeurIPS'24.

PUBLICATIONS AND PREPRINTS

(* Equal contribution † Corresponding author)

Calibrating LLMs with Semantic-level Reward.
Fengfei Yu*, Ruijia Niu*, Dongxia Wu, Yian Ma, Rose Yu†.
arXiv preprint arXiv:2605.15588, 2026.

Generative Model Inversion Through the Lens of the Manifold Hypothesis.
Xiong Peng, Bo Han†, **Fengfei Yu**, Tongliang Liu, Feng Liu, Mingyuan Zhou.
Advances in Neural Information Processing Systems (NeurIPS), 2025.

Model Inversion Attacks: A Survey of Approaches and Countermeasures.
Zhanke Zhou*, Jianing Zhu*, **Fengfei Yu***, Xuan Li, Xiong Peng, Tongliang Liu, Bo Han†.
arXiv preprint arXiv:2411.10023, 2024.

SCHOLARSHIPS & AWARDS

Graduate Student Researcher (Research Assistantship)	2026
Outstanding ASMPT Award	2025
HKSAR Reaching Out Award (ROA)	2025
EDPS Innovative Scholarship	2025
Zih Chi Wen Memorial Scholarship	2025
Study Abroad Programme Scholarship (McGill University)	2024
Vincent Woo Scholarship for Outstanding Mainland Students	2024
Undergraduate Research Programme Scholarship (UGRP)	2024
Concentration Award in Computer Science	2024
Outstanding Student Scholarship	2023
Summer Undergraduate Research Fellowship (SURF)	2023
Undergraduate Scholarship in Computer Science	2022
President's Honour Roll & Dean's List	2021–2025

TEACHING EXPERIENCE

Student Assistant, COMP2016 – Database Management Jan 2024 – May 2024
Hong Kong Baptist University, Kowloon, Hong Kong Part-time
· Supported 100+ sophomore CS students with SQL queries and relational-database project work.
· Provided one-on-one coding support during weekly lab sessions.

Student Assistant, COMP1005 – Essence of Computing Jan 2023 – May 2025
Hong Kong Baptist University, Kowloon, Hong Kong Part-time
· Guided 60+ freshman students through lecture content, coding logic, and debugging across four semesters.
· Mentored students through the design, implementation, and testing phases of their final projects.

CERTIFICATIONS & TEST SCORES

Mathematical Contest in Modeling (MCM) – S Award
IELTS – Overall Band **7.5**

SKILLS

Programming Languages	Python, Java, Kotlin, SQL
ML / DL Frameworks	PyTorch, TensorFlow
Systems & Tooling	Linux, Bash, Conda, CUDA, Git, LaTeX