

Xiaoyu (Nicholas) Wu

Contact	xiaoyuwu@andrew.cmu.edu · 412-510-5091 · Pittsburgh, PA nicholas0228.github.io · Google Scholar	
Research Focus	Generative modeling, representation learning, and reasoning Efficient training and inference; privacy and trustworthy machine learning	
Education	Carnegie Mellon University Ph.D. student in Societal Computing, School of Computer Science Software and Societal Systems Department Advisors: Zhiwei Steven Wu, Niloofar Miresghallah, and Andrew Ilyas	08/26 - 09/30 (expected)
	Rice University Ph.D. studies in Computer Science Research with Chen Wei	09/25 - 06/26
	Shanghai Jiao Tong University M.Sc. in Computer Science	09/22 - 03/25
	Shanghai Jiao Tong University B.S. in Computer Science and Technology	09/18 - 06/22
Publications	Full publication list : Google Scholar · * Equal contribution	
	Recent preprints	
	Compress to Remember: Learning Compact Memory via On-Policy Distillation for Long Video Generation Xiaoyu Wu [*] , Weihang Guo [*] , Yifei Wang [*] , Xinze Feng, Lydia E. Kavraki, Zhiwei Steven Wu. arXiv preprint, 2026 · Paper Train a memory compressor via on-policy distillation (OPD), pairing the same frozen generator with full history as teacher and compressed memory as student.	
	Representation by Design in Generation: Cross-View Class-Token Alignment in Diffusion Transformers Xiaoyu Wu, Yifei Wang, Chen Wei. arXiv preprint, 2026 · Paper Align class tokens across independently noised views, improving ImageNet class-token linear-probe accuracy by 9.4 percentage points while maintaining comparable generation quality.	
	Taming Outlier Tokens in Diffusion Transformers Xiaoyu Wu [*] , Yifei Wang [*] , Tsu-Jui Fu, Liang-Chieh Chen, Zhe Gan, Chen Wei. arXiv preprint, 2026 · Paper Apply Dual-Stage Registers to representation encoders and diffusion denoisers to reduce outlier-token artifacts and improve image generation quality.	
	Visualizing Distribution Coverage in Generative Diffusion Models Yifei Wang, Xiaoyu Wu, Tsu-Jui Fu, Chen Chen, Liang-Chieh Chen, Zhe Gan, Chen Wei. arXiv preprint, 2026 · Paper Use pass@k to reveal that distribution-matching distillation trades distribution coverage for single-sample quality in image and video generation.	
	Scaling Video Generation for Reasoning: At What Cost? Weihang Guo, Xiaoyu Wu, Yifei Wang, Niloofar Miresghallah, Lydia E. Kavraki. arXiv preprint, 2026 · Paper Quantify the compute cost of hidden-state reasoning with a controlled Rubik's Cube benchmark, showing that generation loss alone does not reliably predict reasoning accuracy.	

Publications
(continued)

First- and shared-first-author publications

Exploring Diffusion Models' Corruption Stage in Few-Shot Fine-tuning and Mitigating with Bayesian Neural Networks

Xiaoyu Wu^{*}, Jiaru Zhang^{*}, Yang Hua, Bohan Lyu, Hao Wang, Tao Song, Haibing Guan.
[KDD 2026](#) · [Paper](#)

Identify a corruption stage in few-shot diffusion fine-tuning and mitigate it with Bayesian neural networks, improving generation quality and diversity.

Unlearned but Not Forgotten: Data Extraction after Exact Unlearning in LLM

Xiaoyu Wu, Yifei Pang, Terrance Liu, Zhiwei Steven Wu.
[NeurIPS 2025](#) · [Paper](#)

Extract removed training data using guidance from pre- and post-unlearning logits, exposing privacy risks even after exact unlearning.

Leveraging Model Guidance to Extract Training Data from Personalized Diffusion Models

Xiaoyu Wu, Jiaru Zhang, Zhiwei Steven Wu.
[ICML 2025](#) · [Paper](#)

Extract fine-tuning images by extrapolating between pretrained and personalized diffusion models, then clustering generated samples to identify likely training data.

CGI-DM: Digital Copyright Authentication for Diffusion Models via Contrasting Gradient Inversion

Xiaoyu Wu, Yang Hua, Chumeng Liang, Jiaru Zhang, Hao Wang, Tao Song, Haibing Guan.
[CVPR 2024](#) · [Paper](#)

Assess whether artwork was used for fine-tuning by reconstructing masked details from differences between pretrained and fine-tuned diffusion models.

Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples

Chumeng Liang^{*}, Xiaoyu Wu^{*}, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, Haibing Guan.
[ICML 2023](#) | [Oral](#) · [Paper](#)

Optimize adversarial image perturbations with Monte Carlo sampling to disrupt artistic feature extraction and hinder unauthorized imitation by diffusion models.

Additional publications

Accelerating Inference of Discrete Autoregressive Normalizing Flows by Selective Jacobi Decoding

Jiaru Zhang, Juanwu Lu, Xiaoyu Wu, Ziran Wang, Ruqi Zhang. [TMLR 2026](#) | [Featured Certification](#) · [Paper](#)

Stealthy Backdoor Attack in Federated Learning via Adaptive Layer-wise Gradient Alignment

Qingqian Yang, Peishen Yan, Xiaoyu Wu, Jiaru Zhang, Tao Song, Yang Hua, Hao Wang, Liangliang Wang, Haibing Guan. [ICCV 2025](#) · [Paper](#)

Targeted Attack Improves Protection against Unauthorized Diffusion Customization

Boyang Zheng^{*}, Chumeng Liang^{*}, Xiaoyu Wu. [ICLR 2025](#) | [Spotlight](#) · [Paper](#)

Toward Effective Protection against Diffusion-Based Mimicry through Score Distillation

Haotian Xue, Chumeng Liang^{*}, Xiaoyu Wu^{*}, Yongxin Chen. [ICLR 2024](#) · [Paper](#)

Awards

MIDST Challenge, SaTML 2025 - Winning team, all four tracks.

Membership inference on diffusion-based synthetic tabular data.

[Challenge results](#) · [Related work: TPDP 2025 workshop](#)

Open Source
Impact

Mist - Founder and main developer.

1,000+ combined GitHub stars across [Mist](#) and [Mist-v2](#); featured in [The Verge](#).

Social engagement: **16K reposts** and **20K likes**.

Academic
Service

Reviewer: ECCV, ICML, NeurIPS, ICLR.