DIE CHEN

EDUCATION

National University of Singapore *Singapore*
Ph.D. in Computer Science | School of Computing Aug 2026 – Present
Research Focus: Agent Safety, Trustworthy AI.

East China Normal University *Shanghai, China*
Master of Engineering in Big Data Technology and Engineering | Recommended Admission Sept 2023 – Jun 2026
Research Focus: AIGC, Content Safety, Controllable Generation.

Zhejiang Gongshang University *Hangzhou, China*
Bachelor of Engineering in Software Engineering | Rank: 1/99. Sept 2019 – Jun 2023

ACADEMIC PAPERS

Die Chen, Zhiwen Li, Mingyuan Fan, Cen Chen, Wenmeng Zhou, Yanhao Wang, Yaliang Li. Growth Inhibitors for Suppressing Inappropriate Image Concepts in Diffusion Models. **ICLR 2025 Poster.**

Die Chen, Zhiwen Li, Cen Chen, Yuexiang Xie, Xiaodan Li, Jinyan Ye, Yingda Chen, Yaliang Li. Comprehensive Assessment and Analysis for NSFW Content Erasure in Text-to-Image Diffusion models. **NeurIPS 2025 Datasets and Benchmarks Track Poster.**

Die Chen, Zhongjie Duan, Zhiwen Li, Cen Chen, Daoyuan Chen, Yaliang Li, Yingda Chen. AttriCtrl: A Generalizable Framework for Controlling Semantic Attribute Intensity in Diffusion Models. **ICLR 2026 Poster.**

Zhiwen Li, **Die Chen**, Mingyuan Fan, Cen Chen, Yaliang Li, Yanhao Wang, Wenmeng Zhou. Responsible Diffusion Models via Constraining Text Embeddings within Safe Regions. **WWW 2025 Oral. Best Student Paper Award.**

Zhiwen Li, Zhongjie Duan, **Die Chen**, Cen Chen, Daoyuan Chen, Yaliang Li, Yingda Chen. AutoLoRA: Automatic LoRA Retrieval and Fine-Grained Gated Fusion for Text-to-Image Generation. Manuscript in preparation.

RESEARCH EXPERIENCE

Identify and Correct Unsafe Image Features in the Latent Space During Inference
The 13th International Conference on Learning Representations | 1st Author Mar 2024 – Oct 2024

- **Problem:** Diffusion models learn unsafe content like NSFW and copyrighted material. Keyword filtering fails to block harmful images from ambiguous prompts, while direct fine-tuning risks catastrophic forgetting.
- **Content:** Developed a concept-aware image editing method that identifies unsafe regions based on target NSFW concepts and applies pixel-level correction to minimally modify images while preserving overall quality. It leverages attention mechanisms during inference, avoiding repeated fine-tuning caused by changes in target concepts and catastrophic forgetting.
- **Contribution:** Independently designed and implemented a novel concept erasure algorithm, **achieving SOTA results with 20% higher effectiveness** while preserving image quality and semantics.

Responsible Diffusion Models via Constraining Text Embeddings within Safe Regions
The 34th ACM World Wide Web Conference | 2nd Author Mar 2024 – Oct 2024

- **Content:** Proposed and implemented a novel self-discovery approach to prevent the generation of NSFW and socially biased content. The method identifies a semantic direction vector in the embedding space to constrain text embeddings within a safe region, eliminating the need for explicit keyword filtering or full model fine-tuning. The approach exhibits robust and versatile capabilities, delivering outstanding results in both safe content generation and fair generation scenarios.
- **Contribution:** Designed a lightweight fine-tuning approach and contributed to the codebase design. Conducted experiments focused on NSFW content generation. The work received the **Best Student Paper award.**

Benchmarking NSFW Erasure in Diffusion Models
The 39th Annual Conference on Neural Information Processing Systems | 1st Author Oct 2024 – Jan 2025

- **Content:** Established the first benchmark for unsafe concept erasure in text-to-image diffusion models. Developed a comprehensive end-to-end toolkit tailored for concept erasure evaluation and analysis. Conducted multi-dimensional assessments including cross-method comparisons and intra-method condition analyses.
- **Contributions:** Provided in-depth explanations of the underlying mechanisms influencing baseline performance. Delivered practical recommendations on method selection and application under varying constraints of time, data efficiency, safety, and image quality.

PROFESSIONAL EXPERIENCE

Alibaba Group – Tongyi Lab *Hangzhou, China*
Research Intern | Controllable Text-to-Image Generation Mar 2025 - Sept 2025

- **Research Contributions:** Identified the limitation of conventional encoders in handling discrete attributes, which prevents continuous control over attribute intensity. Proposed AttriCtrl, a universal and plug-and-play control framework featuring a semantic intensity encoder that decouples multiple attributes and enables both independent and joint control to meet flexible user generation demands.
- **Other Contributions:** Integrated AttriCtrl into the open-source framework Diffsynth (10K+ stars on GitHub). Contributed to algorithm development for the Bailian platform and provided technical support for cloud services.

RESERACH PROJECTS

AI-Powered Children's Storybook Platform Apr 2024 – Aug 2024

- **Content:** Developed a voice-interactive platform enabling children to create customizable storybooks. Implemented cross-image attention mechanisms for scene continuity and protagonist embeddings for character consistency, innovatively achieving the restoration of a large number of details of multiple characters.
- **Contribution:** Led team, responsible for frontend architecture and core image consistency algorithm design. Won National 2nd Prize in C4-AIGC Competition (Top 3%) and Shanghai "Tianyi Cloud Cup" 2nd Prize.

AWARDS AND HONORS

Academic Excellence
National Scholarship 2025
Provincial Outstanding Graduate Award 2026, 2023
Provincial Government Scholarship 2020, 2019
The First Prize Scholarship (Top 3%) 2020, 2019
Merit Student (Top 3%) 2020, 2019

Competitions
National 2nd Prize, China Collegiate Computing Contest - AIGC Innovation (Top 3%, Team Leader) 2024
National 2nd Prize, China Collegiate Service Outsourcing Competition (Top 3%, Team Leader) 2022
2nd Prize ,"Tianyi Cloud Cup" Shanghai College Students Computing Application Competition (Team Leader) 2024

Research & Innovation
Best Student Paper award on the ACM on Web Conference 2025 2025
National Undergraduate Innovation and Entrepreneurship Training Program Project (Core Member) 2023
Zhejiang College Students Science and Technology Innovation Event Program (Team Leader) 2022

Patents & Software Copyrights
Invention Patent: AR-Based Personalized Advertisement System (First Inventor, No. CN 114900722 B) 2024
Computer Software Copyright: Video AR Personalized Advertisement Platform (First Copyright Holder) 2022
Computer Software Copyright: MSTick Attendance Management System (First Copyright Holder) 2022