Tao **Huang**

PH.D. CANDIDATE IN COMPUTER SCIENCE

School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China

(+86) 153-1309-8801 | thuang@bjtu.edu.cn | ht86159.github.io | github.com/HT86159 | Tao Huang

Education

Beijing Jiaotong University (*BJTU*) *Beijing, China*

PH.D. CANDIDATE IN COMPUTER SCIENCE AND TECHNOLOGY *Sept. 2023 – Present*

- Advisor: Prof. Liping Jing
- Co-advised by: Dr. Rui Wang (Lecturer)
- Research Focus: Trustworthy AI (Uncertainty Learning & Adversarial Learning)

Beijing Jiaotong University (*BJTU*) *Beijing, China*

B.S. IN MATHEMATICS AND APPLIED MATHEMATICS *Sept. 2019 – June 2023*

- GPA: 3.77/4.0 (Rank: 4/15)

Publications

Visual Hallucination Detection in Large Vision-Language Models via Evidential Conflict

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING (IJAR 2025, JCR Q2, CCF B) *Accepted*

- Authors: **Tao Huang**, Zhekun Liu, Rui Wang, Yang Zhang, Liping Jing
- Accompanying dataset PRE-HAL released on Hugging Face (**13,000+** downloads within 3 months.).

Detecting Misbehaviors of Large Vision-Language Models by Evidential Uncertainty Quantification

SUBMITTED TO INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR 2026) *Under Review*

- Authors: **Tao Huang**, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, Liping Jing

Learning Robust Adversarial Simplex for Black-Box Attack

SUBMITTED TO INTERNATIONAL JOURNAL OF COMPUTER VISION (IJCV) *Under Review*

- Authors: **Tao Huang**, Rui Wang, Zhengwei Fang, Yi Li, Lei Shi, Yi Qin, Tong Xiao, Xiaofei Liu, Huafeng Liu, Michael K. Ng, Liping Jing

Learning a Universal Perturbation with Flat Simplex to Jailbreak Large Vision–Language Models

SUBMITTED TO IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2026) *Under Review*

- Authors: **Tao Huang**, Rui Wang, Xiaofei Liu, Xiaomeng Li, Yi Qin, Liping Jing

SAM Encoder Breach by Adversarial Simplicial Complex Triggers Downstream Model Failures

IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV 2025) *Accepted*

- Authors: Yi Qin, Rui Wang, **Tao Huang**, Tong Xiao, Liping Jing

Strong Transferable Adversarial Attacks via Ensembled Asymptotically Normal Distribution Learning

IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2024) *Accepted*

- Authors: Zhengwei Fang, Rui Wang, **Tao Huang**, Liping Jing

Boosting Adversarial Transferability Across Diverse Contexts via Efficient Implicit Ensemble

IN PREPARATION FOR IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS) *In Preparation*

- Authors: Yi Li, Rui Wang, Yao Zhu, Zhengwei Fang, **Tao Huang**, Yi Qin, Zhekun Liu, Tengkun Yang, Liping Jing

Research Experience

Project 1: Uncertainty Quantification & Learning Based on Evidence Theory *BJTU*

SUPERVISORS: PROF. LIPING JING & DR. RUI WANG *Sep. 2024 – Present*

- **Uncertainty Quantification (Inference Phase):** Proposed a post-hoc quantification framework based on Dempster-Shafer Theory. By modeling evidence weights from output feature vectors, the system rigorously distinguishes between conflict (model contradiction) and ignorance (data absence), effectively detecting anomalies like hallucinations and jailbreaking attempts. This research yielded two academic publications. The findings include an accepted article in IJAR and a submission currently under review at ICLR 2026.
- **Uncertainty Learning (Training Phase):** Designed an evidence-based training mechanism incorporating a weight learning module and a rectified loss function. This drives the neural network to actively capture evidence conflicts and ignorances during feature learning, endowing the model with uncertainty awareness (outputting ”unknown” or ”conflicted” signals) rather than relying on post-calibration. Ongoing efforts aim for a NeurIPS 2026 manuscript.
- **Open Dataset Release:** Developed and publicly released the dataset PRE-HAL on Hugging Face, downloaded **13,000+** times within 3 months.

Project 2: Adversarial Transferability based on Optimization & Flatness *BJTU*

SUPERVISORS: PROF. LIPING JING & DR. RUI WANG *Sep. 2022 – Present*

- **Adversarial Attack Optimization:** Formulated adversarial attack as a constrained optimization problem, leveraging the insight that flat optima correlate with better generalization (transferability). Introduced simplex optimization to learn a flat solution space and employed model filtering to smooth the loss landscape, stabilizing the optimization process and significantly boosting attack transferability. This effort yielded a manuscript presently undergoing review at IJCV.
- **Universal Jailbreak (LUP):** Developed a single perturbation strategy to jailbreak large vision-language models. This approach utilizes simplex optimization for generating perturbations that trigger specific text sequences across diverse datasets. This universal perturbation effectively forces the model to generate predefined text across diverse samples, outperforming complex multi-modal perturbation methods. This effort yielded a manuscript presently undergoing review at CVPR 2026.

Project 3: Misinformation Detection via Multi-Agent Systems *BJTU*

SUPERVISORS: PROF. LIPING JING *Jan. 2024 – Present*

- **Multi-Agent System for Multimodal Fact-Checking:** Constructed a multi-agent system comprising specialized roles (e.g., Text Verifier, Image Verifier, Judge) to validate multimodal information. Simulated interaction dynamics through debate and voting mechanisms, leveraging group consensus to detect deep semantic conflicts and identify complex fake news patterns more accurately than single-agent models.

Mathematical Modeling Contest (MCM/ICM) *BJTU*

TEAM LEADER & MAIN MODELER *Feb. 2022*

- Developed evaluation models using Principal Component Analysis (PCA) for talent assessment and the Analytic Hierarchy Process (AHP) for R&D evaluation, and established a Bi-objective Optimization Model based on a proposed cost function.
- Solved the model using Gradient Descent algorithms, achieving the **Finalist (Top 0.3%)** award.

Honors & Awards

INTERNATIONAL COMPETITION

2022	**Finalist (Top 0.3%)**, MCM/ICM (Mathematical Contest in Modeling)	*International*

SCHOLARSHIPS & DOMESTIC HONORS

2023	**Research Innovation Scholarship**, Beijing Jiaotong University	*Beijing, China*
2023, 2024	**Second Class Academic Scholarship**, Beijing Jiaotong University (Ph.D.)	*Beijing, China*
2021	**Second Prize (Beijing District)**, CUMCM (China Undergraduate Mathematical Contest in Modeling)	*Beijing, China*
2020, 2021	**Second Class Academic Scholarship**, Beijing Jiaotong University (Undergraduate)	*Beijing, China*
2023, 2024	**Outstanding League Member**, Beijing Jiaotong University	*Beijing, China*

Skills

Programming	Python, MATLAB, R
Deep Learning	PyTorch, Scikit-learn, AutoAttack, WandB (Weights & Biases)
Data Analysis	NumPy, Pandas, Matplotlib, Seaborn
General Tools	Git, Linux/Shell, LaTeX
Languages	Chinese (Native), English (CET-6: 605/710)

Academic Service

2024	**Conference Reviewer**, ACM MM 2024	*International*
2024, 2025	**External Reviewer**, CVPR 2024, ICCV 2025	*International*
2023, 2024	**Volunteer**, NSFC Distinguished Young Scholars Evaluation Meeting (Life & Earth Sciences)	*Beijing, China*
Fall 2023	**Teaching Assistant**, Machine Learning (Prof. Liping Jing)	*BJTU*
2023 – Present	**System Administrator**, Laboratory Server Maintenance & Team Event Organizer	*BJTU*