# Sitong Fang

sitongfang1@gmail.com · sitongfang.com · Sitong Fang

*Research Interests: Trustworthy multimodal AI, physically grounded world models, and AI alignment.*

## Education

**Peking University** B.S., 2023 – 2027 (expected)
Yuanpei College, Artificial Intelligence

- **Honors:** Yuanpei Young Scholar; Boya Scholarship; Soong Ching Ling Future Scholarship; Beijing Natural Science Foundation *QiYan* Research Program; Freshman Scholarship (First Prize)
- **Technical Skills:** Python, PyTorch, vLLM, DeepSpeed, Ray, CUDA, C/C++, LaTeX

## Publications

**When Slower Isn't Truer: Inverse Scaling Law of Truthfulness in Multimodal Reasoning** **ACL 2026**
Sitong Fang, Wenjing Cao, Jiahao Li, Xuyao Wang, Chi-Min Chan, Sirui Han, Juntao Dai, Yike Guo, Yaodong Yang, Jiaming Ji

- Discovered an inverse scaling law: slower reasoning models are less truthful in multimodal settings.
- Proposed **TruthfulVQA**, the first benchmark for multimodal truthfulness evaluation with 5,000+ images, and **TruthfulJudge**, a human-in-the-loop evaluation framework.

**Debate with Images: Detecting Deceptive Behaviors in Multimodal LLMs** **Under Review**
Sitong Fang, Shiyi Hou, Kaile Wang, Boyuan Chen, Donghai Hong, Jiayi Zhou, Juntao Dai, Yaodong Yang, Jiaming Ji

- Introduced **MM-DeceptionBench**, the first benchmark for evaluating deceptive behaviors in multimodal LLMs.
- Proposed **Debate with Images**, a multi-agent debate framework requiring models to ground claims in visual evidence, significantly improving deception detection accuracy.

**AI Deception: Risks, Dynamics, and Controls** **Under Review**
Boyuan Chen*, Sitong Fang*, Jiaming Ji*, . . . , Yaodong Yang†, Tiejun Huang†, Ya-Qin Zhang†, HongJiang Zhang†, Andrew Yao† * Equal contribution † Corresponding author

- The first systematic international report on AI deception, with Turing Award laureate Andrew Yao as the corresponding author. Submitted to ACM Computing Surveys.
- Formally defined AI deception using signaling theory, analyzed the deception cycle, and proposed mitigation strategies.

## Research Experience

**Physis AI** Feb 2026 – Present
Research Scientist

- World model startup building physically grounded world foundation models with reinforcement learning. Backed by Hillhouse Ventures and Yanyuan Capital (over $10M first round).

**PKU-Alignment Group, Peking University** Dec 2024 – Present
Research Intern · Advised by Prof. Yaodong Yang

- Led the development of **TruthfulVQA**, the first benchmark for multimodal truthfulness evaluation (ACL 2026).
- Led the development of **Debate with Images**, a visually grounded multi-agent debate framework for multimodal deception detection, along with **MM-DeceptionBench**, the first benchmark for multimodal deception.
- Co-first authored the **AI Deception Survey**, with Turing Award laureate Andrew Yao as corresponding author.
- Core contributor to **Align-Anything** (4,600+ stars on GitHub), an all-modality alignment framework.
- Core contributor to **Eval-Anything**, a comprehensive safety evaluation framework for multimodal models.

**HKGAI R&D Center / HKUST** Dec 2024 – Mar 2025
Research Intern

- Contributed to **HKGAI-V1**, the Hong Kong government's first locally fine-tuned generative AI model based on DeepSeek, supporting Cantonese, Mandarin, and English with localized safety alignment.

## Honors and Awards

2025 Yuanpei Young Scholar, Peking University *10 annual recipients*
2025 Beijing Natural Science Foundation Undergraduate *QiYan* Research Program *Sole recipient in cohort*
2025 Soong Ching Ling Future Scholarship *10 annual recipients university-wide*
2024 Boya Scholarship, Peking University
2024 CMSC Scholarship, Peking University
2024 Academic Excellence Award, Peking University
2024 Social Service Award, Peking University
2023 Freshman Scholarship (First Prize), Peking University
2023 Ranked 1st in Fujian Province, National College Entrance Exam (Science)