

GUANGYU YANG

+44 7304077646

gy266@cam.ac.uk

linkedin.com/in/guangyu-yang

bruceyg.github.io

Education

University of Cambridge

PhD in Engineering

Machine Intelligence Lab

Oct 2024 – Present

Cambridge, UK

Supervisor: Prof. Bill Byrne

University of Cambridge

Master of Philosophy in Machine Learning and Machine Intelligence

(Speech and Language Processing track)

Oct 2022 – Aug 2023

Cambridge, UK

University of California, San Diego - Overall GPA: 3.91 (Cum Laude)

Bachelor of Science in Computer Science - Major GPA: 3.98

Bachelor of Science in Cognitive Science

with a Specialization in Machine Learning and Neural Computation - Major GPA: 4.0

Sept 2018 – June 2022

La Jolla, CA, USA

Research Interests

LLM/VLM Post-Training and Alignment, Reinforcement Learning, AI Safety, Content Moderation, Retrieval-Augmented Generation, Minimum Bayes Risk Decoding.

Grants & Awards

UKRI AI Research Resource (AIRR) Gateway Award

Project Co-Lead (with Prof. Bill Byrne as Project Lead)

2026

10,000 GPU-hours on Isambard-AI

Lambda Research Grant Program

Lead Investigator

2026

\$5,000 in cloud credits

Cambridge International Scholarship

Cambridge Commonwealth, European and International Trust

2024

Full PhD funding

Publications

* Equal contribution; † Corresponding author.

- **Guangyu Yang**^{*}, Jingbiao Mei^{*}, Mingsheng Sun^{*}, Jinghong Chen, Yingtong Bu, Pengda Qin, Da Chen[†], and Bill Byrne[†]. *Controllable Multi-label Video Safety Detection via Adaptive Tversky Policy Optimization*. **NeurIPS 2026**.
- **Guangyu Yang**, Jinghong Chen, Jingbiao Mei, Weizhe Lin, and Bill Byrne. *Retrieval-Augmented Defense: Adaptive and Controllable Jailbreak Prevention for Large Language Models*. **ACL 2026 (Main)**.
- Jingbiao Mei[†], Jinghong Chen, **Guangyu Yang**, Xinyu Hou, Margaret Li, and Bill Byrne[†]. *According to Me: Long-Term Personalized Referential Memory QA*. **PALM @ NeurIPS 2026 (Workshop)**.
- Jinghong Chen, Jingbiao Mei, **Guangyu Yang**, and Bill Byrne. *BERAG: Bayesian Ensemble Retrieval-Augmented Generation for Knowledge-based Visual Question Answering*. **arXiv preprint, 2026**.
- Chenyang Ma, Kai Lu, **Guangyu Yang**[†], Jiuming Liu, Shitong Xu, Bill Byrne, Ioannis Havoutis, Niki Trigoni, and Andrew Markham. *CycleVLA: Proactive Self-Correcting Vision-Language-Action Models via Subtask Backtracking and Minimum Bayes Risk Decoding*. **arXiv preprint, 2026**.
- Jingbiao Mei, Jinghong Chen, **Guangyu Yang**, Weizhe Lin, and Bill Byrne. *Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection*. **EMNLP 2025 (Main, Oral)**.
- Jinghong Chen, **Guangyu Yang**, Weizhe Lin, Jingbiao Mei, and Bill Byrne. *On Extending Direct Preference Optimization to Accommodate Ties*. **NeurIPS 2025**.
- **Guangyu Yang**, Jinghong Chen, Weizhe Lin, and Bill Byrne. *Direct Preference Optimization for Neural Machine Translation with Minimum Bayes Risk Decoding*. **NAACL 2024 (Main, Short)**.
- **Guangyu Yang**. *Multilingual Models in Neural Machine Translation*. MPhil Dissertation, University of Cambridge, 2023.

Research Experience

University of Cambridge, Machine Intelligence Lab

Oct 2024 – Present

PhD Researcher, supervised by Prof. Bill Byrne

Cambridge, UK

- **Agent Safety:** Developing local safeguards for coding agents against indirect prompt injection, using executable attack scenarios, deterministic outcome verification, and privileged-context self-distillation to predict harmful consequences before tool execution.
- **LLM Jailbreak Defense:** Developed Retrieval-Augmented Defense (RAD), a framework that adapts to emerging jailbreak strategies without retraining and enables controllable safety-utility trade-offs. Reduced PAIR-1 attack score from 71.45 to 20.09 on Qwen2.5-14B (ACL 2026).
- **VLM Safety Alignment:** Developed Adaptive Tversky Policy Optimization (ATPO), a reinforcement learning framework for multi-label video safety detection with controllable precision-recall trade-offs. Achieved a Jaccard Index of 75.44 on SafeWatch-Bench-Real, compared with 40.66 for the base model (NeurIPS 2026).
- **Preference Optimization:** Investigated preference-based alignment methods for language models, including extending Direct Preference Optimization (DPO) to accommodate tied preferences (NeurIPS 2025).

University of Cambridge, Machine Intelligence Lab

Oct 2022 – Aug 2023

MPhil Researcher, supervised by Prof. Bill Byrne

Cambridge, UK

- **Neural Machine Translation:** Developed a preference optimization approach using Minimum Bayes Risk (MBR) decoding to generate training preferences, improving translation quality while retaining efficient standard decoding at inference time (NAACL 2024).

Industry Experience

RedNote (Xiaohongshu)

Apr 2026 – Oct 2026

Multimodal LLM Research Intern

Remote

- Developed distributed SFT and GRPO training pipelines for large vision-language models using verl, Ray, FSDP, and vLLM, supporting multi-GPU optimization and efficient rollout generation.
- Designed global and category-specific adaptive reward mechanisms to control precision-recall trade-offs in video safety classification, contributing to Adaptive Tversky Policy Optimization (ATPO, NeurIPS 2026).
- Released trained VLM safety checkpoints and accompanying research artifacts to support reproducibility and downstream safety applications.

To0 Space

Sep 2023 – Sep 2024

Machine Learning Engineer

Remote

- Fine-tuned Stable Diffusion and Vision Transformer models for high-fidelity architectural rendering and image generation.
- Built automated image labelling and captioning pipelines using vision-language models (VLMs) and large language models (LLMs), and integrated retrieval-augmented generation (RAG), ControlNet, and IP-Adapter for architectural image synthesis.

Teaching Experience

University of Cambridge

2025 – 2026

Teaching Assistant and MPhil Dissertation Co-supervisor

Cambridge, UK

- **Teaching Assistant (MLMI 8: Large Language Model Applications, Lent 2026):** Prepared lecture materials on modern LLM architectures, designed and graded coursework, and conducted student supervision and Q&A sessions.
- **MPhil Dissertation Co-supervisor (2025–2026):** Proposed a research project on jailbreaking LLMs for copyrighted content, conducted weekly supervision meetings to guide research direction and experimental design, and provided feedback on the dissertation manuscript.

University of California, San Diego

Spring 2022

Instructional Assistant, COGS 118A: Supervised Machine Learning Algorithms

La Jolla, CA, USA

- Designed and graded homework and exam questions, and provided student support on supervised machine learning algorithms.

Technical Skills

Programming & Systems: Python, C++, SQL, Bash, Linux

ML Frameworks: PyTorch, Hugging Face Transformers, verl

Distributed Training & Inference: FSDP, Ray, vLLM

LLM/VLM Post-Training: SFT, LoRA, DPO, GRPO, reward design, knowledge distillation