Department of Data Science
香港城市大學
City University of Hong Kong

DS SEMINAR

Toward Trustworthy Foundation AI: Systems That See, Know, and Think as We Do

Date: 5 February 2026 (Thursday)

Time: 9:30am - 10:30am

Venue: Rm 6-209, Lau Ming Wai Academic Building, City University of Hong Kong

ABSTRACT

Recent foundation models exhibit a "Paradox of Competence": they can achieve silver medals in the International Math Olympiad yet frequently fail at trivial tasks like counting objects or interpreting basic charts. This disparity reveals that current "black-box" models often rely on statistical shortcuts rather than genuine understanding, making them brittle and difficult to trust in high-stakes collaboration.

In this talk, I present a research roadmap for bridging this gap to build **Trustworthy Intelligence**: systems that are perceptually grounded, reasoning-justified, and architecturally modular. The talk is structured around three key inquiries:

1. **See As We See**: I will first diagnose the "Grounding Gap" using Chimera, a semiotic benchmark that reveals how Vision-Language Models (VLMs) frequently cheat via visual, linguistic, and knowledge shortcuts rather than processing the actual visual signal.
2. **Know As We Know**: I will demonstrate that possessing knowledge is insufficient if it cannot be flexibly applied. Through work on Visual Math and Modality Fusion, we show that models suffer from a "Compositional Bottleneck", failing to integrate basic knowledge/skills such as robust visual recognition with symbolic reasoning.
3. **Think As We Think**: Finally, I will present a technical deep-dive into the mechanisms of **In-Context Learning (ICL)**. By testing the Structured Task Hypothesis, we provide empirical evidence that Large Models naturally generalize by composing latent primitives. I will argue that leveraging this latent compositionality towards explicit **Modular Architectures** is the key to creating AI partners that are not only powerful but transparent, verifiable, and safe.

Mr. Yifan HOU

GUEST SPEAKER'S PROFILE

Mr. Yifan HOU is a final-year Ph.D. candidate in the Department of Computer Science at ETH Zurich, advised by Prof. Mrinmaya Sachan and Prof. Antoine Bosselut. His research centers on the intersection of vision-language understanding, interpretability, and reasoning, with the goal of building AI systems that perceive the world accurately and reason in human-aligned, transparent ways. His work has been published in top conferences including ICML, ACL, EMNLP, and NeurIPS. Previously, he has conducted research at Meta AI, EPFL, and TTIC.

Enquiries: ds.go@cityu.edu.hk

All are welcome