Xuan Luo

6582 Calle Koral, Goleta, CA 93117 | xuan_luo@ucsb.edu | https://luoxuan-cs.github.io/

Research Focus

My research primarily focuses on Efficient Large Language Models:

- Computational efficient model inference via adaptive layer skipping (FlexiDepth, DiffSkip).
- Directly decoding multiple tokens without relying on speculative decoding (DMTD).

Additionally, I also worked on application areas of generative models, including: Detecting ChatGPT imposters with a single question (FLAIR); Image style transfer via progressive alignment (PAMA).

Education

University of California Santa Barbara, CA, USA — Sep 2023 – ongoing
Ph.D. Computer Science
Advisor: Prof. Xifeng Yan

Xi'an Jiaotong University, Shaanxi, China — Aug 2019 – June 2023
B.E. Computer Science (Honor Program)

Publications

[1] Direct Multi-Token Decoding — 2025
Xuan Luo, Weizhi Wang, Xifeng Yan
Preprint
[PDF] [CODE] [WEBSITE]

[2] Adaptive Layer-skipping in Pre-trained LLMs — 2025
Xuan Luo, Weizhi Wang, Xifeng Yan
COLM 2025
Oral Presentation (top 24/418)
[PDF] [CODE] [DATASET] [WEBSITE]

[3] DiffSkip: Differential Layer Skipping in Large Language Models — 2025
Xuan Luo, Weizhi Wang, Xifeng Yan
ACL Findings 2025
[PDF] [CODE]

[4] Bot or Human? Detecting ChatGPT Imposters with A Single Question — 2024
Hong Wang, **Xuan Luo**, Weizhi Wang, Xifeng Yan
COLM 2024
[PDF] [CODE] [DATASET]

[5] Progressive Attentional Manifold Alignment for Arbitrary Style Transfer — 2022
Xuan Luo, Zhen Han, Linkang Yang
ACCV 2022
[PDF] [CODE]

[6] Computer Science Diagram Understanding with Topology Parsing — 2022
Shaowei Wang, Lingling Zhang, **Xuan Luo**, Yi Yang, Xin Hu, Tao Qin, Jun Liu
TKDD 2022
[PDF] [CODE]